

Philosophy of Bioinformatics: Extended Cognition,  
Analogies and Mechanisms

by

Joseph Mikhael

A thesis  
presented to the University of Waterloo  
in fulfillment of the  
thesis requirement for the degree of  
Doctor of Philosophy  
in  
Philosophy

Waterloo, Ontario, Canada, 2007

©Joseph Mikhael 2007

## Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners. I understand that my thesis may be made electronically available to the public.

## Abstract

The development of bioinformatics as an influential biological field should interest philosophers of biology and philosophers of science in general. Bioinformatics contributes significantly to the development of biological knowledge using a variety of scientific methods. Particular tools used by bioinformaticists, such as BLAST, phylogenetic tree creation software, and DNA microarrays, will be shown to utilize the scientific methods of extended cognition, analogical reasoning, and representations of mechanisms. Extended cognition is found in bioinformatics through the use of computer databases and algorithms in the representation and development of scientific theories in bioinformatics. Analogical reasoning is found in bioinformatics through particular analogical comparisons that are made between biological sequences and operations. Lastly, scientific theories that are created using certain bioinformatics tools are often representations of mechanisms. These methods are found in other scientific fields, but it will be shown that these methods are expanded in bioinformatics research through the use of computers to make the methods of analogical reasoning and representation of mechanisms more powerful.

## Acknowledgments

I would like to thank first and foremost my thesis supervisor, Prof. Paul Thagard, for his insights, suggestions, dedication, and, most importantly, his friendship. The process of writing a dissertation relies very heavily on a patient and extensive dialogue between the student and supervisor. In this respect and many others, I could not have asked for anyone better.

I would also like to thank my committee members. Your suggestions have helped greatly develop this dissertation. The suggestions from Prof. Dan Brown were especially appreciated, since his expert knowledge of bioinformatics helped remedy many inaccuracies and misunderstandings I had about the field.

Thirdly, I would like to thank the Department of Philosophy at the University of Waterloo, for accepting and funding me for both my Master's and Ph.D. career. It has been an insightful and enjoyable experience, one which I will look upon with fondness for the rest of my life.

A special thanks goes to my good friend, Nancy Davies, who helped edit and give advice for my dissertation even though she also had her own dissertation to write, as well as a beautiful daughter to raise. Your friendship was and always will be appreciated.

Lastly, I have countless thanks for my family. Your love and example were both the drive and the basis that helped me to pursue my Ph.D. and to complete my dissertation.

## Dedication

Although many, including myself, believe that philosophy helps humanity tackle the most pressing and important questions about reality, humanity, and being, we must also never forget that answers to these questions may be in front of us all along.

This thesis is dedicated to the one person who will always remind me of that fact. To my Taita. Until we meet again.

# Table of Contents

## **Chapter 1. Introduction**

1.1 Biology Today	1
1.2 Bioinformatics	1
1.3 Extended Cognition	5
1.4 Analogical Reasoning	7
1.5 Representations of Mechanisms	9
1.6 Epistemic Appraisal	11
1.7 Extended Cognition in Bioinformatics	12
1.8 Analogical Reasoning in Bioinformatics	18
1.9 Mechanisms and Bioinformatics	19
1.10 Previous Philosophical Investigations into Bioinformatics	21
1.11 Upcoming Chapters	22

## **Chapter 2. Methodological Topics**

2.1 Introduction	23
2.2 Extended Cognition	23
2.2.1 Chalmers, Clark and Extended Cognition	24
2.2.2 Hutchins and Distributed Cognition	25
2.2.3 Giere and Distributed Cognition	26
2.2.4 Wilson and Wide Computationalism	27
2.2.5 Thagard and Distributed Computationalism	28
2.2.6 Humphreys and Extended Cognition	29

2.2.7 Criticisms of Extended Cognition	30
2.2.8 Continuing the Extended Cognition Program	32
2.3 Analogical Reasoning	33
2.4 Representations of Mechanisms	41
2.5 Summary	46
<b>Chapter 3. Bioinformatics</b>	
3.1 Introduction	48
3.2 Molecular Foundation of Bioinformatics	49
3.3 Informational Foundation of Bioinformatics	51
3.4 Computational Foundation of Bioinformatics	52
3.5 DNA Sequencing	53
3.6 DNA Domains	55
3.7 The First Public Database	56
3.8 Genome Sequencing	56
3.9 Database Searching Algorithms	58
3.10 Bioinformatics Websites	58
3.11 NCBI	61
3.12 Microarrays	62
3.13 Summary	64
<b>Chapter 4. BLAST Case Study</b>	
4.1 Introduction	66
4.2 What is BLAST?	67
4.3 Sequence Comparison	67

4.4 History of BLAST	70
4.5 Pairwise Sequence Alignments	71
4.6 Biological Sequence Alignments	74
4.7 The BLAST Algorithm	79
4.8 BLAST Search	82
4.9 The BLAST Case	87
4.10 Extended Cognition in BLAST	94
4.11 Analogical Reasoning in BLAST	97
4.12 Epistemic Appraisal of BLAST	101
4.13 Summary	103
 <b>Chapter 5. Phylogenetic Analyses Case Study</b>	
5.1 Introduction	105
5.2 Classification as Definition	106
5.3 Darwin's Theory of Evolution by Natural Selection	108
5.4 The Molecular Turn	111
5.5 Import from Sequence Comparison Algorithms	113
5.6 Multiple Sequence Alignments	117
5.7 Tree Building	122
5.8 Phylogenetic Tree Case: Human Evolution	126
5.9 Extended Cognition in Phylogenetic Tree Creation	128
5.10 Analogical Reasoning in Phylogenetic Tree Creation	132
5.11 Epistemic Appraisal of Phylogenetic Tree Creation	137
5.12 Summary	140



## **Chapter 6. DNA Microarray Case Study**

6.1 Introduction	142
6.2 History of Microarrays	144
6.3 How Microarrays Work	147
6.4 Image and Data Analysis	153
6.5 Creation of Interactomes	158
6.6 Medical Uses of Microarrays	167
6.7 Extended Cognition in Microarrays	169
6.8 Analogical Reasoning in Microarrays	173
6.9 Mechanisms and Microarrays	176
6.10 Epistemic Appraisal of Microarrays	179
6.11 Summary	181

## **Chapter 7. Conclusion**

7.1 Introduction	183
7.2 Extended Cognition and Bioinformatics	183
7.3. Analogical Reasoning and Bioinformatics	185
7.4 Representation of Mechanisms and Bioinformatics	186
7.5 Epistemic Appraisal of Bioinformatics	186
7.6 Future Prospects	187

<b>Reference List</b>	<b>189</b>
-----------------------	------------

# List of Tables

## **Chapter 1. Introduction**

1.1 Krakatoa/asteroid analogy	8, 9
-------------------------------	------

## **Chapter 2. Methodological Topics**

2.1 Mouse/antidepressant analogy	38
----------------------------------	----

## **Chapter 3. Bioinformatics**

3.1 Bioinformatics milestones	48, 49
-------------------------------	--------

## **Chapter 4. BLAST Case Study**

4.1 Extended cognition comparison in BLAST	97
--	----

4.2 Analogical comparison using BLAST	98
---------------------------------------	----

## **Chapter 5. Phylogenetic Analyses Case Study**

5.1 Rudimentary species comparison	116
------------------------------------	-----

5.2 Species Comparison using multiple sequence alignments	121
---	-----

5.3 Extended cognition comparison in phylogenetic tree creation	131
---	-----

5.4 Analogical reasoning in additions to phylogenetic trees	136
---	-----

## **Chapter 6. Microarray Case Study**

6.1 Microarray milestones	144
---------------------------	-----

6.2 Extended cognition comparison in microarray studies	172
---	-----

6.3 Analogical reasoning in microarray studies	174
--	-----

6.4 Analogical reasoning in viral studies	175
---	-----

# List of Illustrations

## **Chapter 1. Introduction**

1.1 Flashlight mechanism	9
1.2 Synapse signal transmission mechanism	10
1.3 Hemoglobin molecule representation	15
1.4 Secondary structure prediction using MLRC	16
1.5 Structure prediction of hemoglobin using MLRC	17
1.6 Gene expression pattern	20
1.7 Malaria infection mechanism	21

## **Chapter 2. Methodological Topics**

2.1 Tree of life	39
2.2 Lever mechanism	41
2.3 Heart mechanism	43
2.4 Glycolysis mechanism	45
2.5 Microarray preliminary to mechanism creation	46

## **Chapter 3. Bioinformatics**

3.1 Alpha helix and beta sheet	50
3.2 PAM Substitution Matrix	52
3.3 Sanger method of DNA sequencing	54
3.4 DNA domain	56
3.5 NCBI main page	60
3.6 <i>E. coli</i> page from Tax Browser	62

3.7 Microarray	64
<b>Chapter 4. BLAST Case Study</b>	
4.1 Growth of genetic databases	69
4.2 Initial sequence array	72
4.3 First row of values in the sequence array	72
4.4 2 <sup>nd</sup> row of values in the sequence array	73
4.5 Completed sequence array	73
4.6 The Universal Genetic Code	77
4.7 The PAM250 array	78
4.8 Blastp input screen	84
4.9 Results from a Blastp search	86
4.10 Inparalogs vs. outparalogs	92
4.11 OrthoDisease schema	93
4.12 OMIM main page	94
<b>Chapter 5. Phylogenetic Analyses Case Study</b>	
5.1 Darwin's tree of life	109
5.2 Molecular clock	113
5.3 ClustalW input page	119
5.4 ClustalW output	120
5.5 Distance-based tree building	124
5.6 Character-based tree building	125
5.7 Tree created using maximum parsimony	126
5.8 Human tree	128

5.9 Tree of Life Web Project main page	134
5.10 Eutheria page	135
<b>Chapter 6. Microarray Case Study</b>	
6.1 Surface of a microarray	148
6.2 The <i>lac</i> operon	149
6.3 Microarrays	151
6.4 Graph of expression levels	154
6.5 Graph of normalized expression levels	156
6.6 Schema of microarray research	157
6.7 Representation of expression levels	160
6.8 Cluster analysis of gene expression	161
6.9 Representation of T-cell expression levels	162
6.10 Representation of T-cell mechanism	163
6.11 GEO main page	165
6.12 GEO output page	166
6.13 Expression of cancer classes	168
6.14 <i>C. elegans</i> interactomes	177

# Chapter 1

## Introduction

**1.1 Biology Today.** Bioinformatics research employs computational techniques to solve problems in molecular biology. The use of these techniques has been central to such major biological discoveries as the complete description of the human genome (International Human Genome Consortium, 2001), as well being used in recent medical discoveries, such as the annotation of SARS (Wang et. al., 2003), which was instrumental in the development of treatments against the virus (Qiu et. al., 2005, Lu et. al., 2005). However, the success of bioinformatics has to do not only with the computational power that presently exists (although this is a large part of the success), but also with the fact that the discoveries and developments made using bioinformatics employ powerful scientific methods that philosophers and cognitive scientists have discussed over the past few decades. Although there may be other methods employed, the ones that I will look at in this thesis are *extended cognition*, *analogical reasoning*, and *representations of biological mechanisms*. Last, an epistemic appraisal of bioinformatics research will be performed using standards set by Goldman (1992) and Thagard (1997). Before giving descriptions of these methods and standards, however, a quick introduction to bioinformatics is necessary.

**1.2 Bioinformatics.** Pevsner (2003) defines bioinformatics as “the use of computer databases and computer algorithms to analyze proteins, genes, and the complete

collection of deoxyribonucleic acid (DNA) that comprises an organism (the genome).” (p. 3) Bioinformatics has become a recognized scientific field only relatively recently, since about the early 1990s. Among the various discoveries made using bioinformatics, the most recognizable are the following: 1) sequencing of the human genome, along with the genomes of many other organisms, from viruses to the mouse, and 2) annotating newly discovered viruses, such as SARS. Along with these widely recognizable discoveries, others that have been significant to the biological world include: 1) the discovery of various gene and protein functions, 2) the creation of accurate phylogenetic trees, 3) the discovery of complex genetic and protein pathways. Biologists were able to make these kinds of discoveries before the development of bioinformatics, such as the discovery of the Krebs cycle, which is a complex protein pathway (Krebs and Johnson, 1937), but bioinformatics has made similar discoveries possible at an unprecedented rate (Stein, 2005). Bioinformatics has thus been called a *high-throughput* science, meaning that it can produce large volumes of data in relatively short periods of time. Similar advances are being made in other scientific fields, such as physics, astronomy, geology, and climate science. Thus, bioinformatics is no different from these fields with respect to the amount of data produced, but will be shown to be different from other biological fields with respect to the methods used.

Bioinformatics begins with data that are collected using biotechnology instruments, such as genome sequencers, protein crystallization techniques and various gene expression tests. These data are often converted into various standard formats that are recognizable by computer programs. For example, gene sequences are identified with labels like gi|62632718, which is the gene identification number (gi) used by molecular

biologists and bioinformaticists for human embryonic hemoglobin. Using the standardized data, computer algorithms can be run to find relationships among the data and to quantify those relationships.

One example of a typical bioinformatics algorithm that is used for sequence analysis is the Smith-Waterman algorithm (Smith and Waterman, 1981), which tests for local alignments among pairs of sequences. Local alignments are sequence similarities between sequences. One popular program that uses the Smith-Waterman algorithm is the Basic Local Alignment Search Tool (BLAST), which finds sequences that are similar to an input sequence, as well as calculates the probability of similarity under a model of unrelated sequences. The input sequence can be DNA or protein, and the returned sequences are alignments of any of those three types as well (and not necessarily of the type that was input). By comparing sequences, researchers can determine possible functions for newly identified sequences since the stored sequences that are similar may share some functions. For example, if a researcher were to input a protein sequence and the majority of the top matches were retinal-binding proteins, then the researcher can be somewhat confident that the input sequence was also a retinal-binding protein, although many other bioinformatics and laboratory tests can be run in order to help confirm this hypothesis.

Another frequently used set of algorithms compares levels of gene expression. Testing gene expression is important for a number of different reasons. The first is that differences between species may be due to differences in the *expression* of particular genes that they possess rather than *what* genes they possess, thus helping researchers determine the essential differences between species. For example, McConkey (2002)



believes that the majority of genetic differences among humans and chimpanzees may be due to changes in their genetic expression. This difference between expression and composition will be explained in chapter 6. The second is in helping determine which genes are important in differentiating one organ from another. For example, one question a researcher may have is: which genes differentiate the function of brain cells from other cells in the human body? The researcher would then compare the gene expression levels in the brain to other organs, like the liver, kidney and eye. Third, a researcher can compare gene expression levels during the developments of organisms or even specific organs. Lastly, gene expression algorithms can identify differences in gene expression in organs that are normal to ones that are mutated or diseased. These latter comparisons are certainly useful for medical researchers in helping them find cures for diseases and genetic afflictions. One major development in testing gene expression, as well as in biology in general, is the use of DNA microarrays (Schena et. al., 1995), which will be discussed in chapter 6.

A third major set of bioinformatics algorithms are those that create phylogenetic trees based on genomic or protein sequences collected from various organisms. Some of these algorithms use multiple sequence alignment algorithms. Once the sequences are aligned, sequences are grouped based in their relative similarities. These trees are important for biological research in determining the common ancestry of organisms, as well as providing useful insights on evolutionary trends (Thonton and DeSalle, 2000). These modern computational techniques of creating phylogenetic trees are drastically changing phylogenetic trees that were created using morphological, physiological and behavioral comparisons (Hall, 2001). The field of tracing human ancestry has also

blossomed due to the use of these computational techniques. Popular algorithms used in creating phylogenetic trees are PAUP (Phylogeny Analysis Using Parsimony) (Swofford, 1991) and MrBayes (Huelsenbeck & Ronquist, 2001).

Although we will not be studying the Internet in detail, its importance will be apparent while discussing each of the tools described above. Websites such as the one for the National Center for Biotechnology Information (<http://www.ncbi.nlm.nih.gov/>) act as databases as well as allowing users to run a variety of algorithms on the sequences found in the database. This website allows researchers from anywhere in the world to deposit and download sequences, so they can run any algorithm on them. The Internet has allowed researchers to do major biological research with only a few desktop computers connected to the World Wide Web.

There are certainly many other algorithms, tools and websites that are important in bioinformatics research. However, this thesis will limit its investigation to the three described above. With this short description of bioinformatics, we can now do a quick review of several scientific methods that I will later show are relevant to bioinformatics research.

**1.3 Extended Cognition.** In this thesis, ‘extended cognition’ is defined as the accomplishment of a cognitive task using a complex cognitive system involving more than one cognitive individuals and/or representations external to the individuals. Ed Hutchins’ *Cognition in the Wild* (1995) was probably the most famous account of this type of cognition, although he used the term ‘distributed cognition’. He described the phenomenon of ‘pilotage’ on naval crafts and how it is an instance of distributed

cognition. Pilotage is navigation near land, especially when coming into port. Hutchins uses examples such as those of large, military vessels, where a number of sailors using various instruments accomplish pilotage. In this complex cognitive system, no individual person could possibly accomplish this task in the time required to transmit important pieces of information.

Ronald Giere has also devoted many papers to the study of distributed cognition, especially with respect to scientific cognition. He asserts that the scientific revolution in the 17<sup>th</sup> century was not due to a particular change in how people thought about particular problems, but rather was caused by a widespread practice of creating and manipulating *external representations*, such as written symbols. Giere (2003) uses the following example: “Try to multiply two three digit numbers, say 456 x 789, in your head. Few people can do even this very simple arithmetical operation in their heads. Here is how many of us learned to do it:

$$\begin{array}{r} 456 \\ \underline{789} \\ 4104 \\ 3648 \\ \underline{3192} \\ 359784 \end{array} \text{” (pp.2-3)}$$

In this example, external representations are created and manipulated in ways that would be difficult to do in one’s head. Giere therefore added to Hutchins’ definition in that distributed cognition can also involve the manipulation of external representations.

These last cases, I argue, are both examples of extended cognition. The first involves numerous agents or types of agents performing different tasks in order to solve a problem, whereas the second involves the use of external numerous different

representations in order to solve a problem. A more in depth definition of ‘extended cognition’ will be elaborated in Chapter 2.

**1.4 Analogical Reasoning.** The use of analogical reasoning in science has been widely documented by various philosophers as playing a major role in the discovery, development and evaluation of scientific theories. Initially, it was believed that analogies were only important in the discovery and development of scientific theories. Although the logical positivists realized that analogies could be useful in helping a scientist discover and develop new scientific theories, they did not believe that analogies could be used in the evaluation of scientific theories.

Since the work of the logical-positivists, other philosophers have argued that analogies could be used not only in the discovery and development of scientific theories, but also in the evaluation of some of the greatest scientific theories. Famous examples include the analogy between sound and water waves, as well as the analogy between genes on a chromosome and beads on a string. Analogies help scientists visualize the mechanisms that make up their theoretical constructs, as well as help in developing experiments to test particular predictions. Holyoak and Thagard (1995), as well as other philosophers working on analogies, also demonstrated that analogies have been essential in the evaluation of various scientific theories. For example, Darwin used the phenomenon of artificial breeding as evidence that natural selection was an evolutionary force. This thesis will further show that analogies are important for the discovery, development and evaluation of theories by showing their importance in bioinformatics.

Philosophers and psychologists have developed theories of how analogies factor into the discovery, development and evaluation of scientific theories. I will be use the multiconstraint theory of analogy from Holyoak and Thagard (1995) and Shelley (2002). Analogical thinking starts with finding the source and target of the analogy, where the target is the phenomenon to be explained and the source the phenomenon that suggests an explanation. For example, in the water/sound wave analogy, sound was hypothesized to be composed of waves similar to water waves. In this example, the source of the analogy is water waves and the target is sound waves. Next, the elements of both the source and target are compared, and each element is paired with its corresponding analog. For comparison, the elements can be grouped into three categories in the table: the attributes of the analogy, the simple relations of each analog, and the causal relations of the analogs. Shelley (2002) analyzes the following example comparing the debris from a volcanic explosion to the debris from a meteor impact (Alvarez et. al., 1980) using the multiconstraint theory schema:

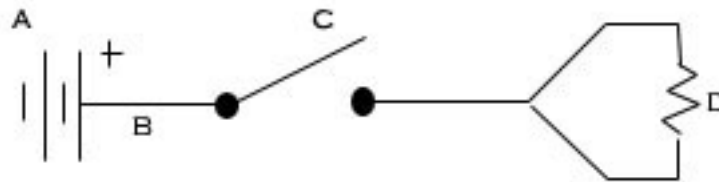
**Table 1.1.** The Krakatoa-asteroid analogy (from Shelley, 2002, p.486).

<b>Krakatoa</b>	<b>Asteroid</b>
Krakatoa-eruption	Asteroid-impact
Debris <sub>k</sub>	Debris <sub>a</sub>
Winds	Winds
Stratosphere	Stratosphere
Earth	Earth
Sunlight	Sunlight
Shade	Darkness
Two-years	Three-years
Eject <sub>k</sub> (Krakatoa-eruption,debris <sub>k</sub> )	Eject <sub>a</sub> (asteroid-impact,debris <sub>a</sub> )
Enter <sub>k</sub> (debris <sub>k</sub> ,stratosphere)	Enter <sub>a</sub> (debris <sub>a</sub> ,stratosphere)
Disperse <sub>k</sub> (winds,debris <sub>k</sub> )	Disperse <sub>a</sub> (winds,debris <sub>a</sub> )
Cover <sub>k</sub> (debris <sub>k</sub> ,Earth)	Cover <sub>a</sub> (debris <sub>a</sub> ,Earth)
Attenuate <sub>k</sub> (debris <sub>k</sub> ,sunlight)	Attenuate <sub>a</sub> (debris <sub>a</sub> ,sunlight)
Persist(shade,two-years)	Persist(darkness,three-years)

Cause <sub>k0</sub> (eject <sub>k</sub> ,enter <sub>k</sub> )	Cause <sub>a0</sub> (eject <sub>a</sub> ,enter <sub>a</sub> )
Enable <sub>k0</sub> (enter <sub>k</sub> ,disperse <sub>k</sub> )	Enable <sub>a0</sub> (enter <sub>a</sub> ,disperse <sub>a</sub> )
Cause <sub>k1</sub> (disperse <sub>k</sub> ,cover <sub>k</sub> )	Cause <sub>a1</sub> (disperse <sub>a</sub> ,cover <sub>a</sub> )
Enable <sub>k1</sub> (cover <sub>k</sub> ,attenuate <sub>k</sub> )	Enable <sub>a1</sub> (cover <sub>a</sub> ,attenuate <sub>a</sub> )
Cause <sub>k2</sub> (attenuate <sub>k</sub> ,persist <sub>k</sub> )	Cause <sub>a2</sub> (attenuate <sub>a</sub> ,persist <sub>a</sub> )

Analogies, along with this schema, will be described in greater detail in the next chapter.

**1.5 Representations of Mechanisms.** Many philosophers have noticed that scientific explanations of phenomena often employ mechanisms. These explanations include a description of the various parts of the mechanism, how they interact with other parts, and the end results or continuing activity of the mechanism. Machamer et. al. (2000) characterize mechanisms as “entities and activities organized such that they are productive of regular changes from start or set-up to finish or termination conditions.” (p. 3) A simple example of this type of explanation is the visual representation of a flashlight, seen in the figure below (Figure 1.1).

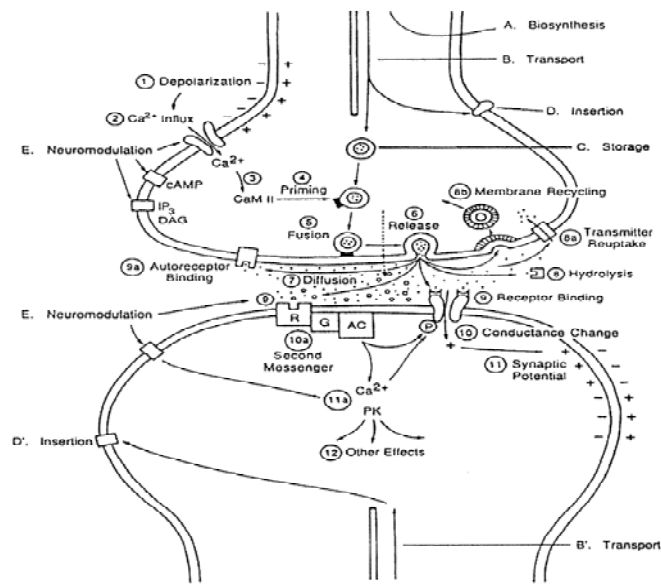


**Figure 1.1.** Diagram of a flashlight, demonstrating a simple machine. A = battery, B = conducting wire, C = switch and D = resistor, or light bulb.

The entities found in this diagram are the four parts: the battery, switch, resistor and light bulb. The activities are the generation of electricity (in the battery), its

conduction (in the wires), its interruption (by the switch) and the resistance of its conduction (by the resistor). A properly working flashlight allows for regular changes from start, (generated electricity flowing through an “on” switch) to its termination (the production of light).

Mechanistic explanations are often well suited for biology, and Machamer et. al. demonstrate this with the working of a synapse (Figure 1.2).



**Figure 1.2.** Visual representation of mechanism of signal transmission across synapses (from Machamer et. al., 2000), with permission.

In this diagram, the various entities, including membranes, protein channels and neurotransmitters, are drawn or named, while the activities and causal links are represented by the arrows between the various entities. The start of the mechanism is the generation of particular signals in one nerve to initiate the release of a neurotransmitter

and the termination of the mechanism is the creation of a signal in another nerve after accepting a neurotransmitter.

A more detailed description of mechanisms and their use in scientific explanations will be given in the next chapter.

**1.6 Epistemic Appraisal.** An epistemic practice is any practice or method that claims to produce knowledge. We will see that although the scientific methods of extended cognition, analogical reasoning and representations of mechanisms are prevalent in bioinformatics research, there is a separate question of how the use of these methods contributes to knowledge. One should be able to judge whether the use of these methods in bioinformatics are reliable and useful epistemic practices when compared to, say, a cognitive task that involves one scientist using the resources of his internal cognition, data from wet-lab experiments, and previous mathematical tools available to the scientist such as statistics.

Alvin Goldman (1992) laid out a set of standards for testing scientific practices. These standards are related to the practices' ability to generate true statements:

1. The *reliability* of a practice is measured by the ratio of truths to total number of beliefs fostered by the practice;
2. The *power* of a practice is measured by its ability to help cognizers find true answers to the questions that interest them;
3. The *fecundity* of a practice is its ability to lead to large numbers of true beliefs for *many* practitioners;
4. The *speed* of a practice is how quickly it leads to true answers.



5. The *efficiency* of a practice is how well it limits the cost of getting true answers.

(Goldman, 1992, p.195)

Thagard (1997) modified and expanded these standards since he believed that they missed important insights about scientific practices. First, scientists are not necessarily in the business of accumulating and testing truths. A better description of their practice is in the collection and evaluation of *results*. Second, results are not only accumulated and evaluated, but scientists also attempt to unify different results with each other. For example, Darwin's work in biology did not just increase the collection of biological facts but unified many facts under one larger theory, namely, the theory of evolution by natural selection. Thus, the following standard should be added to Goldman's list:

6. The *explanatory efficacy* of a practice is how well it contributes to the development of theoretical and experimental results that increase explanatory coherence. (Thagard, 1997, p.255)

I have chosen these standards because they seem to reasonably capture the goals of scientists when they choose one practice over another. These standards will be used when evaluating the success of the use of the aforementioned methods in bioinformatics.

**1.7 Extended Cognition in Bioinformatics.** The first indication that bioinformatics research demonstrates extended cognition is that it requires the work of many different biologists and computer scientists, all of whom have different areas of expertise in this larger field. Bioinformatics research also utilizes different types of external representations, such as the computational representations entered, stored, downloaded,

and manipulated in various computer databases. It is only through the extensive use of computers that bioinformatics is so successful.

Biology already has many sub-fields, from Genetics, to Population Biology, to Ecology, to Developmental Biology. Bioinformatics is a sub-field of Molecular Biology, the latter being concerned with the molecular composition and mechanisms of cells and organisms. However, bioinformatics research is not possible without a great deal of help from computer science.

To give an idea of the extent of computer use in bioinformatics, I will present the following case: Data are initially collected in 'wet labs', laboratories in which biological experiments are performed. These labs collect data such as genome sequences, mRNA expression levels, protein crystallization, DNA knockout tests, as well as many more. The data are then converted to some standardized format to be sent to various computer databases, and thus to be shared by the larger scientific community. For example, genomic sequences are often sent to GenBank (Benson et. al., 2005), a database containing over 100 billion nucleotides collected from 165,000 different species. Information on gene expression levels is often sent to Gene Expression Omnibus (GEO) (Edgar et. al., 2002). Inputted data can be standardized into different formats. Some are stored so that they are easy to read by a human interpreter. For example, TaxBrowser (Wheeler et. al., 2000) contains human readable documentation on species for which there is extensive molecular-biological information. However, other formats, such as FASTA (Altschul et. al., 1990), are formatted so that they are easy for particular computer programs to read. Below is the FASTA format for the human hemoglobin, alpha 1 mRNA sequence.

```
>gi|14456711|ref|NM_000558.3| Homo sapiens hemoglobin, alpha 1 (HBA1),  
mRNA  
ACTCTTCTGGTCCCCACAGACTCAGAGAGAACCCACCATGGTGCTGTCTCCTGCCGACAAGACCAACGTC  
AAGGCCGCCTGGGGTAAGGTCGGCGCGCACGCTGGCGAGTATGGTGCGGAGGCCCTGGAGAGGATGTTCC  
TGTCTTCCCCACCACCAAGACCTACTTCCCGCACTTCGACCTGAGCCACGGCTCTGCCCAGGTTAAGGG  
CCACGGCAAGAAGGTGGCCGACGCGCTGACCAACGCCGTGGCGCACGTGGACGACATGCCCAACGCGCTG  
TCCGCCCTGAGCGACCTGCACGCGCACAAAGCTTCGGGTGGACCCGGTCAACTTCAAGCTCCTAAGCCACT  
GCCTGCTGGTGACCCTGGCCGCCACCTCCCCGCCGAGTTCACCCCTGCGGTGCACGCCTCCCTGGACAA  
GTTCTGGCTTCTGTGAGCACCGTGCTGACCTCCAAATACCGTTAAGCTGGAGCCTCGGTGGCCATGCTT  
CTTGCCCTTGGGCCTCCCCCAGCCCCCTCCTCCCTTCTCTGCACCCGTACCCCGTGGTCTTTGAATAA  
AGTCTGAGTGGGCGGC
```

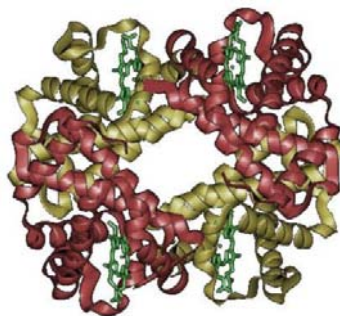
Particular programs such as BLAST can easily read this format so that comparisons between sequences can be quickly made.

The importance of computers is also apparent in the creation of new algorithms to process input data. The Smith-Waterman algorithm has already been discussed for its role in making pairwise comparisons between sequences. Other algorithms include the following: BLAST (Altschul et. al., 1990), which also returns matching sequences from an inputted sequence, and Digital Differential Display, which statistically compares expression levels of genes among organisms, tissues, and cells under differing conditions.

Researchers run these algorithms in order to test various hypotheses. An example of this type of research is when researchers who have sequenced a new protein have no idea of its structure or even its function. They do know, from basic knowledge about molecular biology and physiology, that proteins with a somewhat similar sequence may share the same three-dimensional structure and function. If the researcher uses the NCBI Blast server, the researcher would thus enter his sequence into BLAST and have some similar sequences returned. Many of these returned sequences will have a summary of the function of the protein, and if there is an S next to the returned sequence, then that means that there is also a known structure for that sequence. Thus, instead of taking the

‘wet lab’ route of crystallizing the protein, or determining its function, which requires many expensive and time-consuming experiments where the gene that encodes the protein is knocked out in different ways, the researcher may approximately determine the structure and function in a matter of minutes. However, the results are not as reliable as the wet lab results since the former results are generated using analogical information (more on this in chapters 4, 5 and 6). The researcher can be somewhat confident of these BLAST results, and the wider biological community also shares this confidence as seen by the vast number of publications generated through bioinformatics research. Although BLAST is now used in much more complex programs nowadays (see chapter 4), early uses of BLAST generated some interesting and surprising results (see, for example, Dixon et. al., 1986 and Downward et. al., 1984).

Computer use is also evident in the creation and storage of representations of biological mechanisms. Many biologists use visual representations that appear somewhat similar to the molecules themselves. For example, hemoglobin is often represented in articles as the structure found in figure 1.3.



**Figure 1.3** Visual representation of the hemoglobin molecule (from <http://en.wikipedia.org/wiki/Image:Hemoglobin.jpg>, April 3<sup>rd</sup>, 2006. Permission to copy figure granted by Gnu Free Documentation License).

This representation is very useful for the biologist since the colour version clearly shows the four subunits of hemoglobin, the alpha helices that are characteristic of globin molecules, as well as the heme groups (smaller strands in the middle of each subunit). These heme groups are what carry oxygen molecules from the lungs to the cells and CO<sub>2</sub> molecules back to the lungs. This representation, however, is almost useless for other bioinformatics researchers. Bioinformatics output can be 3-D coordinates, or a probability matrix for types of secondary structures predicted (see figures 1.4 and 1.5). This list of amino acids and their coordinates is almost useless for a biologist, yet is easily readable by many computer programs, such as PDBsum (Laskowski et. al., 1997), which are used to determine the structure and function of the protein. Thus, we see that different types of representation of the same information can be useful in different circumstances. To use Herbert Simon's language (1978), the different representations are 'informationally equivalent' but not 'computationally equivalent'.

```

MLR secondary structure prediction
MPSA code : secondary score MLR
UNK_42590
142 3 HEC
C 0.000000 0.000000 1.000000 M
E 0.020389 0.526834 0.452777 V
C 0.049638 0.377564 0.572798 L
C 0.052460 0.279610 0.667931 S
C 0.112490 0.155684 0.731826 P
C 0.216726 0.055810 0.727464 A
C 0.313164 0.052574 0.634261 D
H 0.584288 0.076941 0.338770 K
H 0.696216 0.068166 0.235618 T
H 0.752292 0.083014 0.164693 N
H 0.735390 0.152764 0.111846 V
H 0.764985 0.148547 0.086469 K

```

**Figure 1.4** Partial output of secondary structure prediction of human hemoglobin, alpha unit using MLRC (Guermeur et. al., 1999).

## Multivariate Linear Regression Combination (SOPMA-GOR4-SIMPA) result

[Abstract](#) Guermeur et al. submitted

View MLRC in: [\[MPSA \(Mac, UNIX\), About...\]](#) [\[AnTheProt \(PC\), Download...\]](#) [\[HELP\]](#)

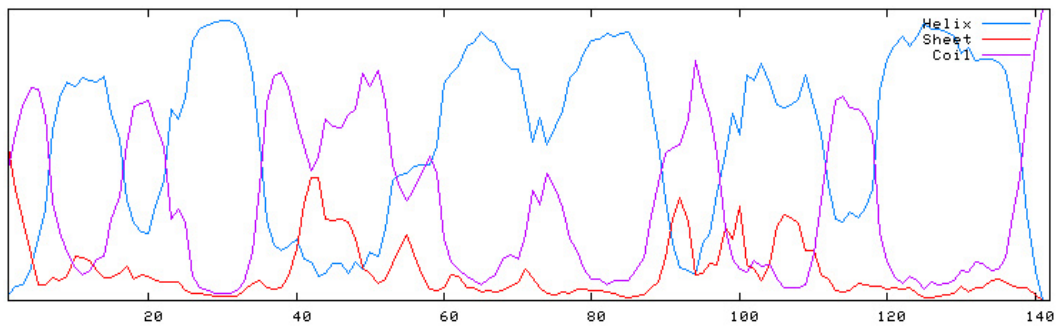
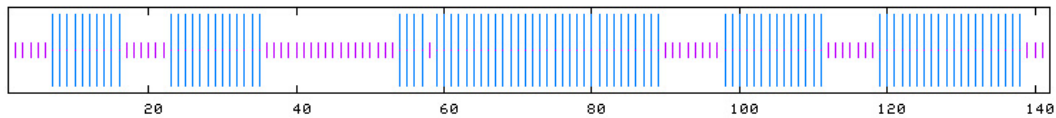
```

      10      20      30      40      50      60      70
      |      |      |      |      |      |      |
MVLSPADKTNVKAAMGKVGGAHAGEYGAELERMFLSFPTTKTYFPHFDLSHGSAQVKGHGKKVADALTNA
ccccchhhhhhhhhccccchhhhhhhhhhhccccccccccccccccccccchhhchhhhhhhhhhh
VAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSK
hhhhhhhhhhhhhhhhhhhhccccchhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhc
YR
cc
  
```

Sequence length : 142

MLRC :

Alpha helix	(Hh)	:	92	is	64.79%
3 <sub>10</sub> helix	(Gg)	:	0	is	0.00%
Pi helix	(Ii)	:	0	is	0.00%
Beta bridge	(Bb)	:	0	is	0.00%
Extended strand	(Ee)	:	1	is	0.70%
Beta turn	(Tt)	:	0	is	0.00%
Bend region	(Ss)	:	0	is	0.00%
Random coil	(Cc)	:	49	is	34.51%
Ambiguous states (?)		:	0	is	0.00%
Other states		:	0	is	0.00%



**Figure 1.5** Structure prediction of human hemoglobin, alpha unit using MLRC (Guermeur et. al., 1999). Output screen from Pole Bioinformatique Lyonnais (Blanchet et. al., 2000), October 3, 2006.

Thus, one very important element that bioinformatics adds to extended cognition is the fact that much of bioinformatics research is done in a ‘dry lab’, or, in

other words, on computers. Where Hutchins (1995) and Giere (2003) concentrate on numerous agents and types of representations, bioinformatics has another element where there are different *types* of agents. Computers have obviously been used for many different tasks, and arguably those tasks are instances of distributed cognition. For example, specialized software is used to find software viruses on computers, and this is a combined cognitive task that is performed by the scanning software and the user. Also, many other academic fields use computers to aid in their research, such as physics, mathematics and architecture. What I believe is unique about bioinformatics, however, is that it is a scientific field that has computer algorithms for doing analogical reasoning.

**1.8 Analogical Reasoning in Bioinformatics.** The three major bioinformatics tools described in the previous section all have one aspect in common: they perform comparative analyses. BLAST compares a query sequence to a database of sequences, be they DNA, mRNA or protein sequences. Microarrays compare the expression levels of cells under varying conditions. Phylogenetic tree algorithms such as PAUP and MrBayes perform multiple sequence alignments, and, through a comparison of those sequences, are able to generate phylogenetic trees. These comparisons form the bases of the analogical reasoning that occurs in bioinformatics. Most instances of analogical reasoning, in both scientific and non-scientific pursuits, involve comparing a target and a source that seem phenomenologically different. For example, the target of these analogies is often some unobservable or unknown entity or mechanism, such as sound waves or asteroid impacts that occurred in the very distant past. This difference presents a potential problem for claiming that analogical reasoning is found in bioinformatics, since both the sources and

targets are often quite similar. But I will try to show that many cases of reasoning found in bioinformatics fit the multiple constraints schema presented by Holyoak and Thagard (1995).

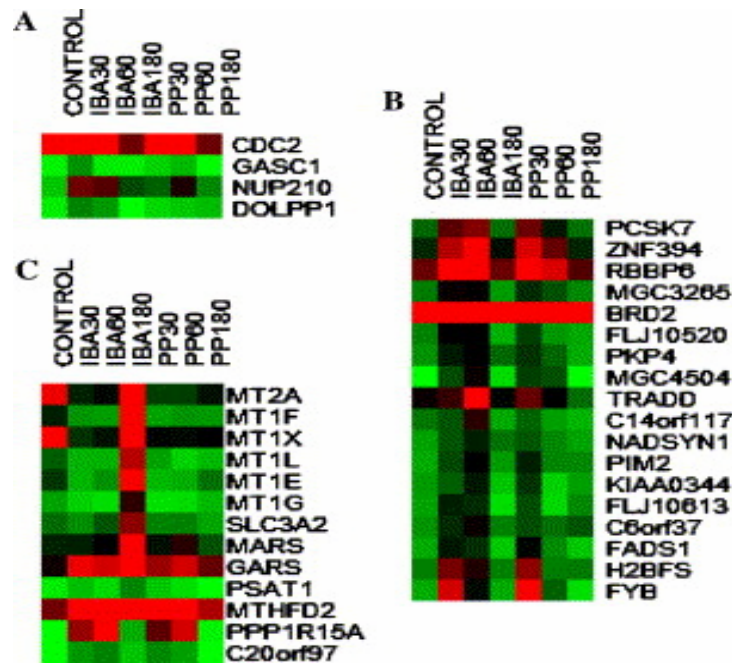
What is striking about the application of analogical reasoning in bioinformatics is that this reasoning is not possible without the aid of sophisticated computer programs, unlike other scientific cases when human minds perform analogies. It will be argued that bioinformatics research uses a unique and powerful version of analogical reasoning due to the use of specific computer algorithms.

**1.9 Mechanisms and Bioinformatics.** With the computational power of bioinformatics, many types of analyses have become possible that were virtually impossible before. For example, analyses of the functions of particular genes or proteins were restricted to being investigated only one or two at a time. The invention of DNA microarrays, however, has enabled researchers to study the actions of as many genes or proteins as they need, including concurrently studying the activities of entire genomes. I have described DNA microarrays in the ‘bioinformatics’ section and will give a more detailed description in chapter 6.

The relationship between microarrays and mechanisms can be seen with the following explanation: Since proteins are hypothesized to be interconnected to bring about the various functions found in a cell, biologists can hypothesize about how they are interconnected by measuring the activity of numerous proteins at once. For example, Yamashita et. al. (2005) studied the change of gene expression in particular blood cells after being infected by viral agents such as those that cause malaria. After infecting



various cells lines (a cell line is a group of cells that are often found clustered together and perform a particular function) with specific viral agents, the researchers found the expression pattern represented in figure 1.6 from the cells after using microarrays.



**Figure 1.6** Expression patterns of genes after infection by malaria (from Yamashita et. al., 2005). Not shown in this diagram are the specific colours of the squares, which are essential in reading expression patterns. The columns represent the cell lines used and the rows represent genes studied.

Using this data, they were able to create the representation of the biological mechanism found in figure 1.7:



aid in scientific discovery. Her summary reviews how many contemporary philosophers of science and scientists recognize science as a problem solving activity, and that heuristics can be created in order to aid in solving those problems.

One specific instance of computational scientific discovery that Darden presents is Karp's paper on the creation of a database on the metabolic pathways of *H. influenza* genome and metabolic pathways. This database was created through the use of analogical reasoning, where the *H. influenza* database was developed using the knowledge base from a similar database for *E. coli*. My thesis will expand on Darden's work by elaborating how computers are essential to bioinformatics, how they improve the methods of analogical reasoning and representations of mechanisms, and how they contribute to biological knowledge.

**1.11 Upcoming Chapters.** This thesis will expand on the ideas presented above. Chapter 2 has detailed descriptions of the methods described above: analogical reasoning; extended cognition; and representations of mechanisms. Chapter 3 gives a detailed historical account of the development of bioinformatics in biology and its importance within that field. Chapters 4, 5, and 6 describe the methods used in particular bioinformatics applications such as sequence alignments using BLAST, the analysis of DNA microarrays and phylogenetic reconstruction. They also appraise these applications using Goldman (1992) and Thagard's (1997) epistemic appraisal standards. Chapter 7 summarizes the scientific methods used in bioinformatics research and suggests other venues for philosophical research into bioinformatics.

## Chapter 2

# Methodological Topics

**2.1 Introduction.** The philosophical investigation of bioinformatics concerns how bioinformatics theories are discovered, developed and evaluated, and how bioinformatics methodology may be different from other scientific disciplines. There are three scientific methods that I have found relevant to bioinformatics. The first is *extended cognition*. My presentation of this method is an extension of similar philosophically analyzed methods, including *distributed cognition*, *distributed computationalism*, and *wide computationalism*. The second method, *analogical reasoning*, has been discussed since at least John Stuart Mill, but has seen a resurgence of discussion with the development of cognitive science. The third scientific method is theorized from the mechanistic view of scientific theories, which is a relatively new view in philosophy of science. This chapter will describe how each applies to bioinformatics. Case studies will be presented in later chapters to further support my preliminary account.

**2.2 Extended Cognition.** Recent discussion in philosophy of science and philosophy of cognitive science has focused on how many aspects of scientific progress have not depended upon particular innovations by single individuals, or on the use of a particular scientific method, but rather by using various external aids. These external aids are important since they extend the existing cognitive abilities found in humans. Edwin Hutchins (1995) and Ronald Giere (2004, 2003, 2002a, 2002b) present the concept of

‘distributed cognition’. David Chalmers & Andy Clark (1998) and Paul Humphreys (2004) present a similar concept called ‘extended cognition’. Robert Wilson (1994) presents a concept called ‘wide computationalism’ and Paul Thagard (1993) presents a similar concept called ‘distributed computing’. I will be reviewing each of these concepts, as well as presenting criticisms. However, I believe that all of these concepts are more similar than different, and closely represent how external aids help human cognizers.

**2.2.1 Chalmers, Clark and Extended Cognition.** In their 1998 paper “The Extended Mind”, Clark and Chalmers presented a thought experiment with the following three cases:

- (1) A person sits in front of a computer screen which displays images of various two-dimensional geometric shapes and is asked to answer questions concerning the potential fit of such shapes into depicted "sockets". To assess fit, the person must mentally rotate the shapes to align them with the sockets.
- (2) A person sits in front of a similar computer screen, but this time can choose either to physically rotate the image on the screen, by pressing a rotate button, or to mentally rotate the image as before. We can also suppose, not unrealistically, that some speed advantage accrues to the physical rotation operation.
- (3) Sometime in the cyberpunk future, a person sits in front of a similar computer screen. This agent, however, has the benefit of a neural implant which can perform the rotation operation as fast as the computer in the previous example. The agent must still choose which internal resource to use (the implant or the

good old fashioned mental rotation), as each resource makes different demands on attention and other concurrent brain activity. (Clark & Chalmers, 1998, pp.10-11)

Clark and Chalmers claimed that cognition had occurred in all three of these cases. Cases 1 and 3 are easy to imagine as cases of cognition because all the operations are happening 'in the head'. However, Clark and Chalmers also argued that the results from cases 1 and 3 are the same as case 2. The only difference in case 2 is that the subject pressed a button in order to achieve the result. Clark and Chalmers argued that the mere pushing of a button, which is an operation that occurs 'outside the skull', does not exclude the operation from being cognitive. Thus, Clark and Chalmers (1998) claimed that cognition does not necessarily have to occur within the mind of an individual. However, they extended their theories by showing that cognition can partly occur in a machine. Thus, as well as allowing cognition to be distributed among individuals and using external representations from fingers, written numbers and models, cognition can be extended to the operations of a computer. Chalmers and Clark called their theory 'extended cognition'.

**2.2.2 Hutchins and Distributed Cognition.** Edwin Hutchins (1995) was probably the first to popularize a notion of extended cognition with the presentation of the concept 'distributed cognition' in his book *Cognition in the Wild*. Referring to an example of pilotage aboard a naval vessel, Hutchins demonstrated how distributed cognition operates. Pilotage is the act of bringing a naval vessel into port, and such a process, especially on larger vessels, requires many operations to be carried out by numerous members of the vessel, as well as the use of various external representations and artifacts.

For example, sailors on each side of the ship telescopically record angular locations of landmarks relative to the ship's gyrocompass. These readings are then passed on to the pilothouse where they are combined by the navigator on a specially designed chart to plot the location of the ship. Once the navigator manipulates the information, the results are passed to the captain who makes the final decisions as to what action the helmsmen should make next. This example shows that the act of pilotage is not carried out by any one individual but requires a number of individuals, and, nowadays, likely the use of computers as well. Also, the information that is used to arrive at the final decision is processed by a number of individuals; therefore, it is not one person who is presented with all the available information from start to finish to make a decision. Hutchins thus makes the following conclusion: The brain is not always the unit of cognition. Cognitive tasks require larger entities such as a whole body, a body plus a tool, or a group of people committed to a particular task.

**2.2.3 Giere and Distributed Cognition.** Ronald Giere (2002a, 2002b, 2003, 2004) applied distributed cognition to scientific progress. Giere (2003) argued that distributed cognition overcomes a dichotomy present in philosophy of science today between cognitive theories of science and social theories of science. Cognitive theories hold that advances in scientific knowledge are dependent upon cognitive attributes that humans possess, such as evaluating coherence between scientific statements and scientific discovery through analogical reasoning (Holyoak & Thagard, 1995). Social theorists, on the other hand, believe that these advances are due to social forces that exist among humans, such as working within a paradigm, or starting scientific revolutions. Distributed

cognition takes both the cognitive and the social spheres into account, allowing for individual human cognitive abilities as well as the interactions between humans to drive scientific advances.

Giere (2002a) argued that one of the most useful tools in scientific research is the scientific model, which can consist of complex graphical mechanisms or mathematical relationships that are predicted among entities. These models are perfect examples of external representations. Lastly, Giere (2002b) argued that one of the causes of the scientific enlightenment from the 16<sup>th</sup> century and onwards was the greater use of external representations such as Cartesian graphs, mathematical models, and animal models.

**2.2.4 Wilson and Wide Computationalism.** In his paper “Wide Computationalism” (1994), Wilson foreshadowed Clark and Chalmers (1998) by also stating that the use of computers challenged any view that saw the human mind as operating in isolation from its environment. Wide computationalism opposes the cognitive scientific theory of internalism, where the latter theory, according to Wilson, is based on the following argument:

1. Cognitive psychology taxonomically individuates mental states and processes only *qua* computational states and processes.
2. The computational states and processes that an individual instantiates supervene on the intrinsic, physical states of that individual.

Therefore



3. Cognitive psychology individuates only states and processes that supervene on the intrinsic, physical states of the individual who instantiates those states and processes. (Wilson, 1994, p.352)

Wilson rejected the internalist thesis because he believed that a computational system is not only restricted to states within an individual but extends to objects in the individual's environment. According to Wilson: "If there are computational descriptions of both an organism's environment and its mental states, and causal transitions from the former to the latter that can be thought of as computations, there is a process beginning in the environment and ending in the organism which can be viewed as a computation, a wide computation." (p.363) Thus, Wilson presented a thesis similar to Chalmers and Clark where cognition is not limited to processes that occur "within the skull" but can include objects and manipulations that occur in one's environment.

**2.2.5 Thagard and Distributed Computationalism.** Thagard (1993) discusses how computer networks are able to solve problems that are too complex for any individual node within the network. Thagard shows how these computer networks are similar to social networks that sociologists such as Latour (1987) have hypothesized to occur in scientific communities. Although not wanting to strictly compare scientific reasoning to a sociological phenomenon, Thagard's project is to show that just as there is an analogy between computers and minds, there is also an analogy between computer networks and scientific communities. Specifically, the nodes of a computer network are analogous to individuals of a scientific community, like collaborators, teachers and students,

colleagues and acquaintances. Just as computer networking has increased the output of computers, scientific networking has increased the output of scientific research.

**2.2.6 Humphreys and Extended Cognition.** In his book *Extending Ourselves* (2004), Paul Humphreys gave an account of the philosophical implications of extended cognition. One of these implications is in supporting a realist view of scientific knowledge. Van Fraassen (1980) argued that entities are real if they are observable by the naked eye, whereas scientifically postulated entities such as planets, molecules, and bacteria that require instruments such as telescopes and microscopes to detect them cannot be considered real, unless there was some way to observe them using the naked eye, such as traveling to the observed planets, or shrinking to the size of an atom. This anti-realist view of scientific entities is supported by the fact that the functions of the tools that are used to observe these entities are dependent upon scientific theories. Thus, if those theories turned out to be false, which is not unlikely since scientific theories are often falsified, then the entities ‘detected’ by those instruments would be called into question.

Humphreys challenged the anti-realist view by arguing that the view confuses the order of reliability between entities that are detected using our unaided senses and entities that are detected using various tools. He argued that we are mistaken more often when making judgments using our unaided senses than when using scientific instruments. For example, our senses can detect the temperature of the air surrounding us, yet instruments that are used in detecting temperature, such as thermometers, are, *ceteris paribus*, much more accurate. Another ability of ours that has been improved upon using computers is our ability to recognize particular humans. We are generally somewhat reliable at

recognizing other people based on their faces and voices, but we often are mistaken, such as when we try to remember the name of a new member of our academic department. However, computers that analyze fingerprints, DNA and even programs that are able to perform facial recognitions have proven to be far more reliable than their human counterparts. This increased reliability is not only apparent within the scientific sphere, but within the legal sphere as well, with DNA evidence becoming more trustworthy in legal cases than eyewitness reports.

As with instruments that augment, convert and extrapolate data that are normally collected by our senses, various instruments have also surpassed our mathematical abilities, especially when one considers the computational power that exists in our current technological world. In *Extending Ourselves*, Humphrey's showed how computers are indispensable in solving complex mathematical problems in modern physics research.

**2.2.7 Criticisms of Extended Cognition.** Robert D. Rupert (2004) presented two major challenges to what he calls the Hypothesis of Extended Cognition (HEC). The first concerns particular consequences we would have to allow if cognition were extended. These consequences include accepting that inanimate, external objects are part of the cognitive process, a process which most of us intuitively accept only happens in our heads. The operations of the brain, according to cognitive scientists, seem to be very different from most external objects that supporters of extended cognition claim to 'extend' our cognition. For example, a note-pad may act as an external memory bank if we jot to-do lists on it. It is still very different, however, from our internal memory since we do not expect our internal memory to look anything like a to-do list on a note-pad.

Another troubling consequence is the assumption in cognitive science that the unit of cognitive research, which is the human individual, would have to be dropped. Cognitive science generally claims to study the functions of individual humans, not groups of humans or a human and some external objects.

Although these are certainly valid objections, Rupert's arguments are based on some misunderstandings of the extended cognition project. The first is that proponents of extended cognition never claimed or required that external objects or processes completely resemble internal ones. There are certainly some differences, yet there are also some similarities, and it is argued by proponents of extended cognition that the similarities outweigh the differences. For example, the contents of my note-pad and those of my memory are certainly different, yet they both have the function of *helping me remember important tasks*. Extended cognition theorists also take advantage of this difference, since they are needed in order to show that these external objects help our internal cognitive processes. To expand on the example, although my memory is somewhat reliable, I often find that if I do not jot down my daily tasks on some kind of paper or in my daily planner, then I end up missing many important meetings or forgetting important tasks since my neural pathways continuously degrade. The reliability of my daily planner is so great due to the relative permanence of ink on paper, and I often feel completely useless if I forget my planner at home.

Also, I do not think that all proponents of extended cognition are claiming that all research in cognitive science includes external objects and processes. Studying internal components and processes is still a very important research project. However, what these proponents are saying is that much of cognition, such as pilotage or modern scientific

research, needs to take into account these external objects and processes in order to fully explain the cognitive steps taken to fulfill those particular tasks.

**2.2.8 Continuing an Extended Cognition Program.** My interest in extended cognition is partly to provide a criticism of the Cognitive View of Scientific Theories. This view holds that scientific theories are mental representations (Thagard 1988, Giere 1988, 1999). More specifically, the main aspects of scientific theories, which include their discovery, development and evaluation, can be understood as cognitive processes, which are computational. I will argue that scientific theories include representations of scientific entities that can be found in computer databases. This thesis will attempt to show, therefore, that many theories in science, especially in bioinformatics, are not only internal mental representations of mechanisms, but can include representations that extend into computer databases, and can be developed using computer algorithms.

In reply to Rupert, I am not claiming that scientific theories must be extended in external objects and processes. The individual scientist can still memorize, discover, develop and evaluate many scientific theories on his or her own, without the aid of any computers or other external objects or processes. What I am attempting to demonstrate, however, is that many contemporary scientific theories are heavily dependent these upon external objects and processes. More specifically, I will show that biological theories that are discovered, developed and evaluated with bioinformatics tools use complex computer algorithms and are contained within large computer databases. The analysis of these biological theories supports an extended cognitive view of scientific theories more than any other view of scientific theories, because no individual scientist can discover,

develop or evaluate these particular biological theories on his or her own, and these theories are hardly ever found outside a massive computer database. It will also be shown that extended cognition helps in making other methods, analogical reasoning and representations of mechanisms, much more epistemically reliable.

I have chosen the concept ‘extended cognition’ not because I think Chalmer’s and Clark’s presentation is superior to that of other authors. I believe all the presentations of some kind of externally aided cognition, including distributed cognition, wide computationalism, and distributed computationalism, are essentially very similar. I believe that the term ‘extended’ is simply most appropriate for describing the phenomenon. This thesis will be supported using particular case studies in chapters 4, 5 and 6.

**2.3 Analogical Reasoning.** Another method that is applicable to bioinformatics research is the use of *analogical reasoning* in the generation, development and evaluation of scientific theories. This topic has been prevalent throughout the history of philosophy of science. Aristotle wrote about analogies, or *shared abstractions* (Shelley, 2003), yet he never applied it to scientific theorizing. John Stuart Mill was probably the first philosopher to further develop this theory (Brown, 1989). In his *System of Logic* (1873), he defined analogies as inferences that are made about the possible properties of one thing based on the known properties of another with which it shares other properties. This definition of analogical reasoning has generally been accepted in the history of philosophy of science, as we will see when looking at the works of more contemporary philosophers. Despite his interest in analogies, Mill did not think that they always played

a role beyond helping make scientific discoveries. Analogies are "...supposed to be of inductive nature but not amounting to complete induction." (Mill, 1873, p.101). However, if the resemblance between the two things being compared is very great, then an analogy may amount to a complete induction. It does not seem as if Mill gives any measure as to how similar the objects need to be in order to be evaluative, however.

Mary Hesse (1952, 1966) described the use of models that give some representation of the phenomena that a theory is attempting to describe. These models can be based upon analogies between the unknown phenomena and some known phenomenon. Hesse's favorite example was the kinetic theory of gas. In envisioning the workings of the gas molecules and how they are affected by temperature, pressure and volume, one is often given the analogy of billiard balls. Just as billiard balls move and bounce off of one another according to particular forces applied to them, similar actions occur on gas molecules. According to Hesse, the analogy does not only play a positive role in the discovery of the theory, but can also help in evaluating the theory. The analogous relationship between the known and unknown phenomena is three-fold: positive, negative, and neutral. The positive portion of the analogy is what initiated the researcher to propose the analogy in the first place. In the billiard ball/gas molecules analogy, the entities involved are seen as spherical and as obeying the laws of Newtonian mechanics. The negative portions of the analogy are the properties of both phenomena that are assumed not to have any relationship whatsoever. Gas molecules are not expected to be colored, solid or striped, for example, nor are billiard balls free-floating in the air. Lastly, the neutral portion consists of the aspects of the analogy that are not yet known. Just as billiard balls move faster or slower depending upon mechanical forces that are

applied unto them, the kinetic theory was tested as to whether factors such as pressure, temperature, and volume of a container affect gas molecules in a mechanical manner. The neutral aspect of an analogy, therefore, helps in the development of a theory just as the positive aspect helps in its discovery.

Other researchers have developed similar theories on how analogies are used in science (Bonner 1963, Wilson 1964, Lee 1969). Ruse (1973) was probably the first researcher to see the importance of analogies in biological research, especially with respect to Darwin's use of analogies in supporting his theory of evolution by natural selection. The analogies Darwin used, namely comparing natural selection to Malthus' economic theories on the Welfare state, as well as to artificial selection, are among the most famous in the history of science. The presentation by Darwin of these analogies gives the impression that they were not only used to discover and develop the theory of natural selection, but also to serve as evidence for the theory of evolution by natural selection. Darwin's extensive description of artificial selection, how varieties are created through the selections that farmers and breeder make, and how those varieties eventually create new species, give support for the natural selection thesis. The difference between artificial and natural selection is that the former requires selection pressures from a human farmer or breeder whereas the latter requires selection pressures within the environment.

The relationship between science and analogies was brought back into consideration among those who held a cognitive view of scientific theories. As a part of the cognitive view, the use of analogies was seen as a special kind of cognitive tool. Holyoak and Thagard (1995) studied the use of analogies in many spheres of thought,



from reasoning in childhood, to courtroom cases, to political decisions, philosophical discussions, and, of course, scientific theorizing. What these authors also attempted to show, as hinted by the title of their book, *Mental Leaps*, is that analogical reasoning is a very important tool in helping to solve many problems. For example, they presented many cases where individuals are told a particular story where problem solving was involved, and when those individuals were presented with a story that was somewhat similar to the problem, they were much better than a control group at solving the problem. Such mental leaps are not uncommon in many aspects of human cognition, from children's stories to historical comparisons.

Analogies have also figured importantly in many of history's most influential scientific theories. Holyoak and Thagard listed the following analogies that helped in the discovery, development, and even evaluation of scientific theories: Sound/water waves, Earth/small magnet, movements of the Earth/movements of the Moon, movements of the Earth/movement of a ship, light/sound, planetary motion/projectiles, lightning/electricity, respiration/combustion, motive power of heat/motive power of water, animal and plant competition/human population growth, natural selection/artificial selection, electromagnetic forces/continuum mechanics, benzene ring/self-cannibalistic snake, chromosome/beaded string, bacterial mutation/slot machine and mind/computer.

Building on the structure-mapping theme of Gentner (1983), Holyoak and Thagard gave specific details on the relationships between the source of an analogy and its target. These elements are often brought together by some kind of similarity between the elements of the source and the target. Second, there are structural parallels that should exist between the source and the target. These parallels should also be one-to-one,

meaning “each element of the target domain should correspond to just one element in the source domain (and vice versa)” (Holyoak & Thagard, 1995). Lastly, analogies generally have purposes, that is, they are used by an individual to fulfill some task. In the case of scientific theories, analogies are used in order to discover, develop or evaluate some theory. These three details compose the *multiconstraint theory* of analogical reasoning. With these constraints, searching for analogies and deciding whether they are adequately used in supporting scientific theories becomes an easier task.

A similar treatment of analogical reasoning has been offered by Cameron Shelley in a number of works (2002, 2003, 2004). Shelley used tables that map the relationship between a source and target according to various dimensions. The first dimension shows the correspondence between certain objects, properties or attributes of a source and a target. The second dimension has the correspondence between the relations among the attributes. Finally, the third dimension shows the correspondences between the system relations of each domain. Below is an example of this table (Table 2.1) using an analogy Shelley (2004, p.4) constructed between the use of anti-depressants in a mouse and its use in humans:

In the Porsolt Forced-Swim Test (Porsolt et al. 1977), a mouse is placed in a cylinder of water and watched to see how long it swims until it gives up trying to climb out. (The mouse is not drowned at that point but assumes a static position with its hind feet on the cylinder bottom and its nose out of the water.) It turns out that mice that are treated with antidepressants tend to swim longer than normal mice do. This model of the action of antidepressants enjoys construct validity in

the sense that the increased time that mice treated with antidepressants spend trying to extricate themselves from the cylinder corresponds to the increased hope for success in life that depressed people treated with antidepressants feel in the pursuit of their goals. (Shelley, 2004, p.4).

**Table 2.1** Table showing mapping between a source and a target in an analogy.

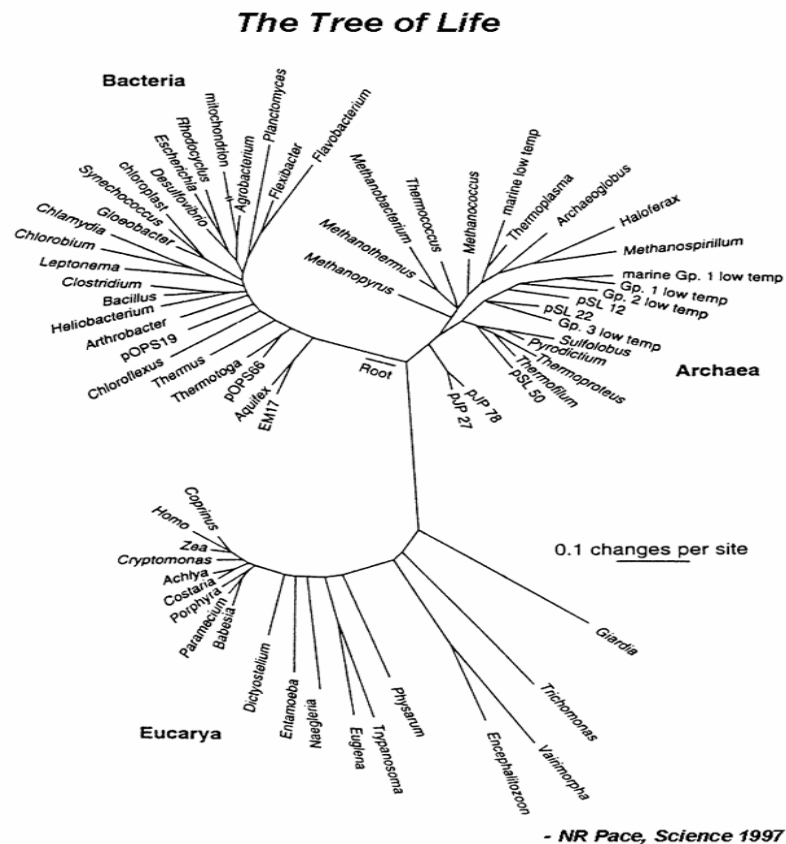
<b>Mouse</b>	<b>Human</b>
Mouse	Patient
Antidepressant <sub>m</sub>	Antidepressant <sub>h</sub>
Safety	Goals
Longer-time	Further-extent
Look-for (mouse, safety)	Hope-for (patient, goals)
Persist <sub>m</sub> (mouse, longer-time)	Persist <sub>h</sub> (patient, further-extent)
Receive <sub>m</sub> (mouse, antidepressant <sub>m</sub> )	Receive <sub>h</sub> (patient, antidepressant <sub>h</sub> )
Because <sub>m</sub> (persist <sub>m</sub> , look-for & receive)	Because <sub>h</sub> (persist <sub>h</sub> , hope-for & receive)

Based on the information given in the Porsolt Forced-Swim Test, and the correspondence between the properties, relations, and system-relations in the description, it is hypothesized that there is a clear analogy between the action of the anti-depressant in mice and its action on humans.

Using the theory of analogical reasoning described so far, especially Shelley's tabular comparison, I will show how much of bioinformatics research is the application of analogical reasoning. The extensive use of analogical reasoning in bioinformatics is due to the fact that much bioinformatics research involves making inferences on unknown sequences and processes based on those that are familiar.

The extensive use of analogical reasoning in bioinformatics is based on findings and assumptions found in evolutionary biology and molecular biology. First,

bioinformatics research is necessarily embedded in evolutionary theory. One of the major claims of evolutionary theory is that all species ultimately descended from one common ancestor. As species evolve, changes accumulate between species. However, many similarities also remain. The accumulation of changes increases with time; therefore, species that are more closely related have fewer changes than those that are more distantly related. This general explanation is often represented using a “Tree of Life”, with many branches of life extending from that original ancestor, and the proximity of the branches representing the evolutionary descent and relationships among species (see figure 2.1).

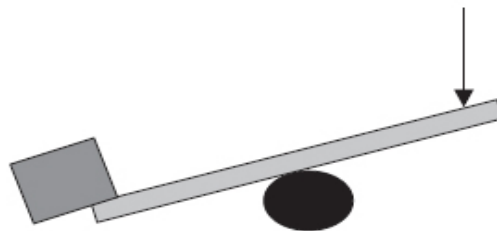


**Figure 2.1** The tree of life. The original ancestor is predicted to be at the ‘root’ of the tree (from <http://www.genex2.dri.edu/research/Tree.gif>).

The evolution of species is also found to be directly correlated with changes in their DNA. The more distantly related two species are, the more changes to the DNA are expected, and vice-versa. If you were to compare two closely related species, their DNA should be relatively similar. Since molecules such as DNA and proteins are compared among species, then analogical reasoning in bioinformatics seems to be a natural progression of previous works in molecular biology. In pre-bioinformatics research, biomolecules researched in non-human species were often thought to be analogical to those found in humans. As such, analogical reasoning is not exclusively used in bioinformatics. However, analogical reasoning is used in unique ways in bioinformatics research. Chapter 4 will look at a tool called BLAST that compares sequences like DNA, RNA and protein among species. These comparisons help in determining relationships between species, discovering the functions of those sequences (assuming that sequences that are closely related are more likely to share the same function), and so on. Many bioinformatics tools are similar to BLAST in making these kinds of comparisons. The quick survey of BLAST in chapter 1 already shows that various comparisons are being made, and often these comparisons are between sequences of known function or descent and sequences of unknown function of descent. This type of comparison fits the description of analogies given above, thus supporting the thesis that analogical reasoning is used in bioinformatics. What is unique about analogical reasoning in bioinformatics, however, is that the method is made even more powerful through the use of computers. Chapters 4, 5 and 6 will further demonstrate the analogical nature of these comparisons, as well as the unique nature of computer use in analogical reasoning found in bioinformatics.

**2.4 Representations of Mechanisms.** In the past decade, *representations of mechanisms* have been found relevant to the nature of scientific models, explanations and methodology (Bechtel and Richardson 1993, Machamer, Darden and Craver 2000, Thagard 2003). Mechanisms, according to Machamer et. al. (2000), are “entities and activities that produce regular changes from start up to termination” (p.3). Bechtel and Abrahamsen (2005) give the following similar definition: “A mechanism is a structure performing a function in virtue of its components parts, component operations, and their organization. The orchestrated functioning of the mechanism is responsible for one or more phenomena.” (p.423). The mechanistic view is particularly applicable to biological explanations, in particular to neuroscientific explanations (Machamer et. al., 2000) and medical explanations (Thagard, 2004). Thagard states that mechanisms do not necessarily require start-up and termination conditions as described by Machamer, but can be involved in various feedback loops.

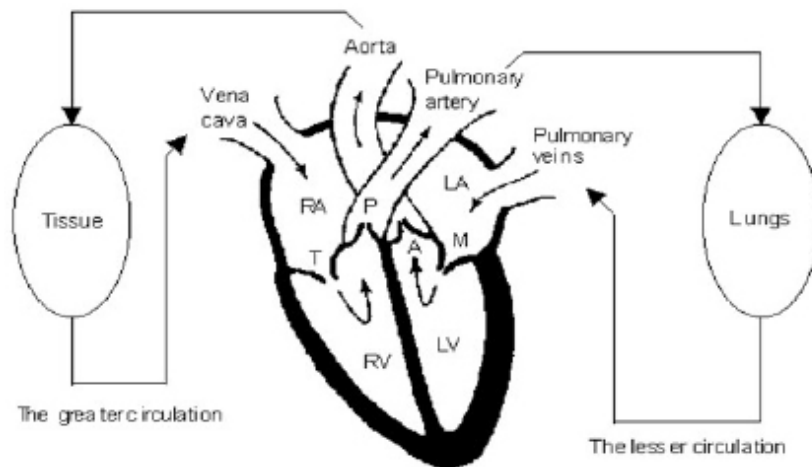
Mechanistic explanations are similar to the descriptions of machines, where there is a description of the various parts, how they interact with other parts, and the end results that are produced by the machine. For example, Thagard gives the following example of a simple machine with very different interacting parts (Figure 2.2).



**Figure 2.2.** Diagram of a lever, demonstrating a simple machine. (from Thagard, 2004, p.55).

The lever has two major parts. The first is the plank and the second is the fulcrum. The plank rests atop the fulcrum and applying force to one side of the plank allows the other side of the plank to rise, thus easily allowing objects that are placed on the other side to be raised to a particular height. The movement of the plank to raise the object constitutes the activity of this mechanism. This simple mechanism can also be part of a larger mechanism, such as a Rube Goldberg machine. Therefore, mechanisms can contain “nested hierarchies” where the part of one mechanism can be explained by referring to the representation of another mechanism (Machamer et. al., 2000, p.13).

Mechanistic explanations are prevalent in biological research, especially in ‘lower-level’ fields such as molecular biology and neurobiology. However they can be found in other scientific disciplines, as well. These disciplines include ecology with the presentation of cycles in ecosystems, and chemistry with the presentation of particular chemical reactions. Thagard (2004) demonstrates the prevalence of mechanistic explanations in medical research involving the study of molecular interactions in a cell. Von Eckardt and Poland (2004) demonstrate the use of mechanistic explanations in cognitive neuroscience. Machamer et. al. (2000) present an example of a biological mechanism through the working of a synapse. The diagram in figure 2.3 from Bechtel and Abrahamsen (2005) is a figure of the human coronary mechanism:



**Figure 2.3.** Visual representation of the human coronary mechanism. From Bechtel and Abrahamsen (2005, p.4).

In this diagram, the various entities, which are the heart membranes, valves and lungs, are either drawn to mimic the actual organ or simply given a structural representation. The activities, indicated by the arrows between the various entities, are the causal links between all the entities. Diagrams, as seen in the above two figures, are visual representations of mechanisms. Two other representations of mechanisms are propositional and schematic. For example, Thagard (2004) gives two mechanisms of virus infection, the first propositional and the second schematic.

*Propositional:* “Viral release may directly cause cell damage or death, as when the SARS virus infects epithelial cells in the lower respiratory tract. Second, the presence of the virus will prompt an autoimmune response in which the body attempts to defend itself against the invading virus: this response can induce symptoms such as high fever that serves to slow virus replication.” (Thagard 2004, p.56)

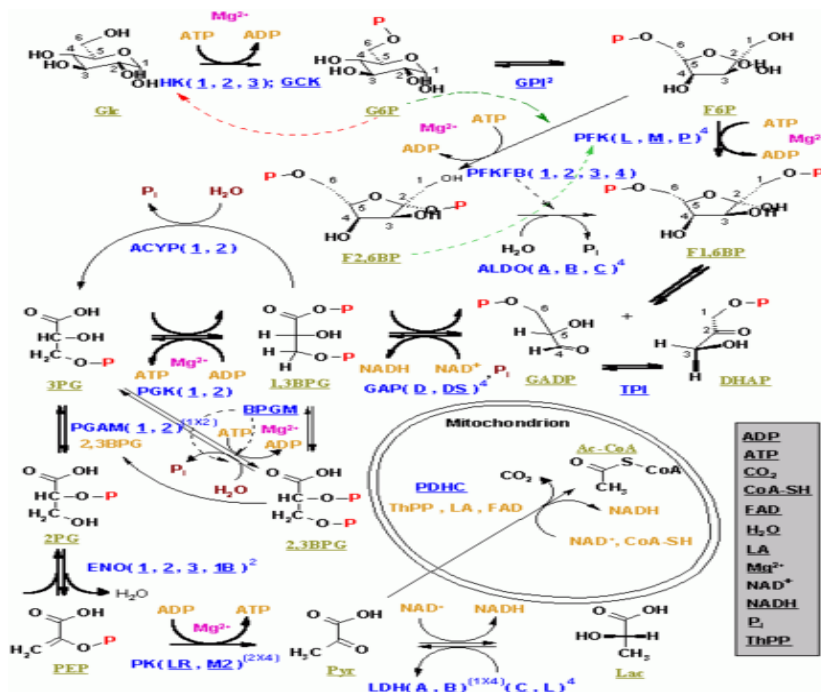


*Schematic:* “viral infection → cell damage → symptoms

viral infection → immune response → symptoms” (Thagard 2004, p.56)

Both representations have each of the components (entities, activities, regular changes) of the mechanistic explanation. Other representations are possible, as it is becoming possible in scientific literature to present mechanisms through 3-dimensional diagrams and even animated video clips. Each representation is useful depending upon the mechanism one is presenting. For example, describing the different neurotransmitters and their passage across a synapse is more difficult to present as a propositional statement. The process of viral infection, although possible to represent in visual form, is easier in a schematic format.

Although these mechanisms are meant to give a representation of the scientific phenomenon, they are not expected to be complete. Within each mechanism there are implicitly embedded mechanisms, and mechanisms are not expected to accurately depict the temporal progression of each step in a mechanism or the spatial dimensions that are involved. Consider the mechanism of glycolysis, which is the mechanism by which glucose is used to create ATP, our bodies’ energy source (figure 2.4):

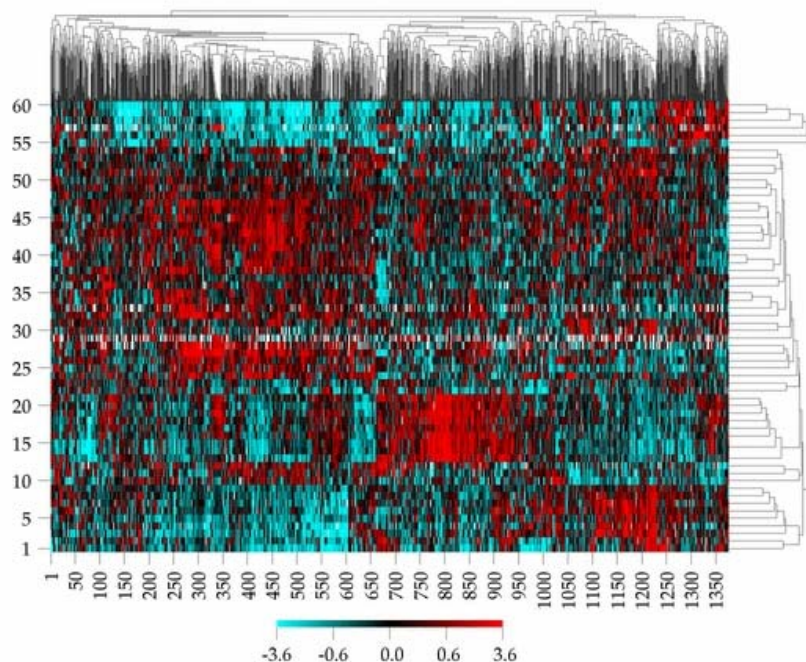


**Figure 2.4** A mechanistic representation of glycolysis. (image is in the public domain).

This mechanism simply gives the entities and activities that are involved in glycolysis, the starting conditions and the termination conditions. There is no indication as to how quickly these reactions occur, where in the cell the reactions occur, nor is there any indication of the proximity of all the entities to each other. These details are not always required in order to effectively explain or model this reaction.

Since bioinformatics tools, such as BLAST and microarrays, are used in analyzing biological molecules, they often produce representations of mechanisms. Many computer programs have been specifically designed to quickly and reliably produce these representations. The most salient of these is the creation of representations based on expression patterns in microarrays (figure 2.5). Using specific programs, the data from this figure can be used to create representations of mechanisms,

a process which would be very difficult for a human scientist. This complete process will be described in greater detail in chapter 6.



**Figure 2.5.** A preliminary representation of a mechanism based on a cell's expression patterns. The branches that are closest to one another represent the molecules that are most likely to interact with one another, whereas those that are furthest are less likely to have a direct interaction. (From Weinstein et. al., 1997).

Chapter 6 will further demonstrate how computers have facilitated the use of representations of mechanisms in bioinformatics.

**2.5 Summary.** This chapter has summarized three important scientific methods that are discussed in philosophy of science today: extended cognition, analogical reasoning and mechanistic representations. Extended cognition is the use of external representations such as computer databases to help solve scientific problems and to store scientific

theories. Analogical reasoning is a process that helps in the discovery, development and evaluation of scientific theories by comparing the previously unknown components of that theory's domain to phenomena that are already known. The *multiconstraint* theory of analogical reasoning, which looks at the similarity, structure and purpose constraints on analogies, will be used in this thesis to demonstrate the use of analogies in bioinformatics. Mechanistic representations describe the operation of entities and activities of a system from start to termination or in a feedback loop.

These methods will be shown in the upcoming chapters to be extensively used in bioinformatics research. It will also be shown that the combination of these methods, i.e. computer use in combination with analogical reasoning and in combination with representations of mechanisms, makes bioinformatics research somewhat unique in its methods. This will be done by presenting three case studies of tools used in bioinformatics, BLAST, phylogenetic studies, and microarrays.

## Chapter 3

# Bioinformatics

**3.1 Introduction.** Although the term was first coined in 1991 (Stein, 2005), developments that are central to bioinformatics were made since the 1960s. It was not until the 1990s, however, that biologists recognized that computational techniques were beginning to make significant contributions to both biological and medical discovery.

The National Center for Biological Information (NCBI) has a list of milestones in bioinformatics, which I have included in table 3.1, along with some additions I have made myself (marked with a ‘\*’). This chapter provides a detailed summary of most of these milestones

**Table 3.1** Bioinformatics Milestones from NCBI ([www.ncbi.nlm.nih.gov/Education/BLASTinfo/milestones.html](http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/milestones.html), my additions marked with a ‘\*’)

<b>Year</b>	<b>Milestone</b>
1962	Pauling's theory of molecular evolution developed
1965	Margaret Dayhoff's Atlas of Protein Sequences compiled
1970	Needleman-Wunsch algorithm developed
1977	DNA sequencing and software to analyze it developed
1981	Smith-Waterman algorithm developed
1981	Concept of the sequence motif developed
1982	Phage lambda genome sequenced
1982	GenBank Release 3 made public
1983	Sequence database searching algorithm developed
1985	FASTP/FASTN: fast sequence similarity searching algorithm developed
1988	National Center for Biotechnology Information (NCBI) created at NIH/NLM

---

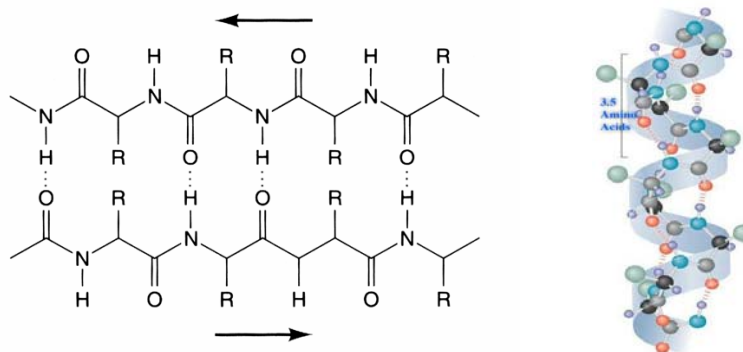
1988	EMBNET network for database distribution developed
1990	BLAST: fast sequence similarity searching algorithm developed
1991	EST: expressed sequence tag sequencing algorithm developed
1993	Sanger Centre created
1994	EMBL European Bioinformatics Institute created
1995	First bacterial genomes completely sequenced
1995*	DNA microarray developed by Affymetrix
1996	Yeast genome completely sequenced
1997	PSI-BLAST algorithm developed
1998	Worm genome completely sequenced
1999	Fruit Fly genome completely sequenced
2004*	Human genome completely sequenced

---

**3.2 Molecular Biological Foundation of Bioinformatics.** One of the greatest scientists of this century, Linus Pauling, is credited as one of the founders of bioinformatics. He is known for his work in quantum chemistry, which is the application of quantum physics to discovering the structure of molecules and their chemical bonds. His work in this field allowed him to develop the technique of X-ray crystallography (Pauling, 1939). This technique shows the structure of molecules by bouncing X-rays off those molecules and studying the pattern the refracted rays make on a plate. Although Pauling designed this technique for use on inorganic molecules, he turned his attention to biological molecules. By using X-ray crystallography on proteins, he was able to show that the secondary structures of proteins, which is the local structure formed by the protein molecules, were mainly composed of alpha helices and beta sheets (figure 3.1).

A brief explanation is necessary at this point to describe the structure of proteins as they occur in living cells. There are four structural levels of proteins. The primary structure of proteins is the actual protein sequence, a simple list of the individual amino acids. The secondary structure is any local structure in the protein sequence caused by the

interactions of the amino acids in the protein, of which alpha helices and beta sheets are examples. The tertiary structure is the overall shape of the protein, while the quaternary structure is the shape a larger protein that is composed of smaller proteins, or sub-units.



**Figure 3.1.** Left: Beta sheet. Most diagrams of proteins with beta sheets represent them as arrows. Right: Alpha helix. Most diagrams of proteins with alpha helices represent them as a helix (Permission to use diagrams granted under the Gnu Free Documentation License).

Pauling used X-ray crystallography to try to determine the structure of DNA. Many believe that Pauling would have been the first to discover its structure if he had access to better equipment or had attended a conference in England where Rosalind Franklin presented high quality diffraction photos of DNA (Lwoff et. al., 1979).

The use of X-ray crystallography on biological molecules was the first technique to give a glimpse into this microscopic world. Since bioinformatics involves the computational investigation of biological molecules, Pauling is seen as the founder of the molecular-biological portion of bioinformatics (Pevsner, 2003).

**3.3 Informational Foundation of Bioinformatics.** As Pauling is the founder of the molecular portion of bioinformatics, Margaret Dayhoff could be considered as the founder of the informational portion. Although work on sequencing proteins was already underway when she presented her work in the 1960s, Dayhoff was the first to develop computer applications that constructed larger sequences from data that had smaller sequences with overlapping peptides. She also created computer programs that were able to accept input from X-ray crystallography experiments. Lastly, Dayhoff was the first to develop computer methods that compared protein sequences, and was the first to derive evolutionary histories of species based on those alignments. This method of finding evolutionary relationships has remained central to bioinformatics. We will see a few of these programs later in the chapter. Dayhoff also compiled the *Atlas of Protein Sequences* (1965), and was the first compilation, or database, of known protein sequences. It contained only 65 proteins, but subsequent volumes made increasing additions to this database.

Dayhoff also created her ‘Substitution Matrix’, which gave the substitution probabilities of amino acids within sequences over evolutionary time. By analyzing the proteins that were available, and by hypothesizing their evolutionary relationship, Dayhoff was able to estimate the probability that particular amino acids were substituted for other amino acids, and the probability that they remained the same. From these probabilities, Dayhoff calculated the *log-odds* ratio, where the score  $S$  for an alignment of amino acid residues  $a$  and  $b$  is given by:

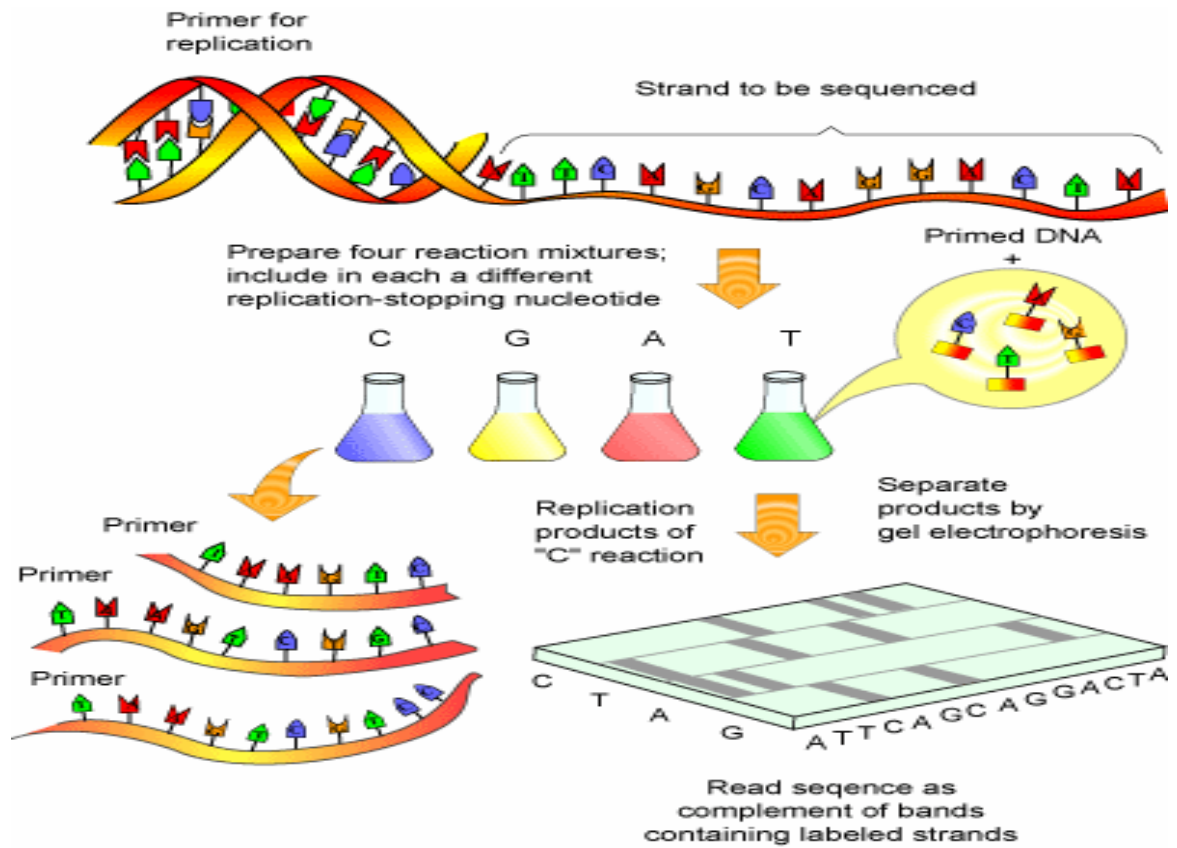
$$S(a, b) = 10 \log_{10}(M_{ab}/p_b) \text{ (from Pevsner, 2003, p.57)}$$





and Smith-Waterman algorithms (1981). These algorithms are designed to compare protein and DNA sequences, thus paving the way to finding evolutionary and functional relationships between sequences. Although sequence-matching algorithms were already being used for other applications, the Needleman-Wunsch and Smith-Waterman algorithms were specifically designed with biological sequences in mind. Algorithms comparing biological sequences are not only designed to find matches, but need to provide scores for degrees of match, since those that are more closely matched are more likely to be evolutionary related to each other. Molecular evolution also causes sequences to have vast areas in a sequence that contain insertions or deletions of DNA or amino acids. These algorithms take these insertions and deletions into account, and are able to modify the resulting score based on this contingency. We will look at these algorithms in greater detail in Chapter 4.

**3.5 DNA Sequencing.** Another important tool that is not necessarily a part of but has developed greatly through collaboration with bioinformatics is DNA sequencing. Although protein sequencing was being performed using mass spectrometry and X-ray crystallography, DNA sequencing was not possible with these techniques due to the large size of DNA molecules. In 1977, Sanger developed a method of sequencing DNA that involved the use of radioactive labels (Lewin, 1997). The Sanger method was ingenious since it was able to create multiple DNA strands from an original strand that were of variable length containing radioactive tags, one for each of the four nucleotides, and the sequence was determined by calculating the lengths of these strands (see figure 3.3).

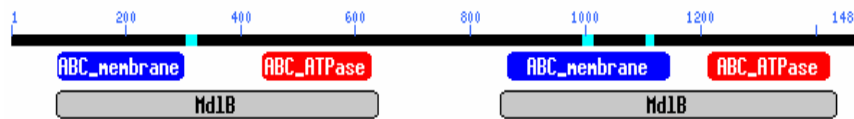


**Figure 3.3.** The Sanger method of DNA sequencing. (From <http://www.bioteach.ubc.ca/Bioinformatics/GenomeProjects/>. Permission for use granted by David Ng and bioteach.ubc.ca.)

In the same year that Sanger developed DNA sequencing, Staden (1977) designed a program that inputs DNA sequences in order to store, edit and analyze such sequences. More modern techniques detect the fluorescence used with a laser and these readings are directly inputted into a computer. Other strategies, such as “shotgun” sequencing, have been employed to make the sequencing process even quicker and cheaper. When the human genome project began in 1998, total sequencing output was 200 Mb for the year. In January 2003, one single institute sequenced 1.5 billion bases for the month

(doegenomes.org). Also, the cost was drastically reduced, from about \$2 per nucleotide at the beginning of the project to about 10 cents per nucleotide by the end of the project (from NCBI genome database at [ncbi.nlm.nih.gov/entrez/query.fcgi?DB=genomeprj](http://ncbi.nlm.nih.gov/entrez/query.fcgi?DB=genomeprj)).

**3.6 DNA Domains.** By 1981, with an increasing number of genes and proteins sequenced and greater analyses into the functions of these molecules, biologists were beginning to recognize that genes and proteins have *domains* (Doolittle, 1981). Domains are areas of a gene or protein that perform a particular function. For example, a transmembrane protein can have three domains. The first is the extracellular component, which is responsible for either importing specific molecules into a cell, exporting specific molecules out of a cell, or both. The second domain is the transmembrane area, which is found in the cell wall, and keeps the protein attached to the cell. The last domain is the intracellular component, which has the same function as the extracellular component, but in reverse. The importance of this discovery to bioinformatics is that it adds extra relevance when searching for similarities between sequences. Nowadays, when bioinformatics researchers perform web-based searches of similar genes or proteins with a query sequence, they first get a page describing the domains that are present in that gene or protein (figure 3.4). The importance of these domains will also be demonstrated in chapter 5 when we look at how phylogenetic trees are constructed using bioinformatics tools.



**Figure 3.4.** The domains of the transmembrane protein responsible for conductance regulation in cystic fibrosis in humans. This figure is returned after performing a sequence similarity search using the web-based NCBI BLAST program. One can click on these motifs and a page will be displayed describing the properties and functions of the motifs.

**3.7 The First Public Database.** Since Dayhoff's *Atlas of Protein Sequences* was released, other databases of gene and protein sequences began to be developed. GenBank is one of those databases (Benson et. al., 2005), and is still extensively used today (you can access the bank and make gene submissions via <http://www.ncbi.nlm.nih.gov/Genbank/index.html>). Its third wide release was in 1982, and this release was significant since it was the first that was publicly available to the scientific community. This public availability started a trend in gene and protein sequencing, and will figure into the argument I will later make about the extended nature of bioinformatics.

**3.8 Genome Sequencing.** With the increasing size of databases and the ability to share information, the ability to sequence whole genomes became easier. It was not long before the first genome, that of the phage lambda virus, was fully sequenced (Sanger et. al, 1982). The phage lambda has a genome only 49 kb in length (or 48,502 nucleotides,

which is relatively small compared to the human genome, which is 3.2 Gb in length). The phage lambda is a bacteriophage that infects *Escherichia coli*; it is probably a good candidate for genome sequencing because not only is its genome very small, it is also easy to replicate, which would provide many samples to sequence. With the complete sequence of a genome, biologists were becoming more confident that they would be able to get a full understanding of how living organisms work, since they had access to their DNA blueprints.

The complete sequencing of genomes followed a trend of increasing organism complexity. The first bacterial genome to be sequenced was the *Haemophilus influenzae* genome, which was 1.83 Mb in size. *H. influenzae* is a bacterium responsible for causing ear and respiratory infections, as well as meningitis in children (Fleischmann et. al., 1995). The first sequenced eukaryotic genome was *Saccharomyces cerevisiae*, or baker's yeast (Galibert et. al., 1996), with a size of 12 Mb. The flatworm, or *Caenorhabditis elegans*, was the first multi-cellular genome to be sequenced (*C. elegans* Sequencing Consortium, 1998) with a genome has 97 million base-pairs. The first 'complex' organism to be sequenced was the fruit-fly, *Drosophila melanogaster*, which was and continues to be a classic test-subject used by geneticists (Ridley, 1996), and with the sequenced genome, researchers finally had access to its 120 Mb of genetic information (Adams et. al., 2000). Currently, the genomes of about 180 species have been fully sequenced, mostly from bacteria and archea, but also including those of the mosquito, honeybee, dog, bread mold, rat, mouse, pufferfish, chimpanzee, and, of course, humans. A complete list of sequenced genomes can be found at the following NCBI web page: <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?DB=genomeprj>.

**3.9 Database Searching Algorithms.** As protein databanks grew in size, it became apparent that algorithms were necessary in order not only to compare individual sequences, as was done with the Needleman-Wunsch and Smith-Waterman algorithms, but also to compare a query sequence with entire databases. This algorithm also had to be somewhat efficient yet without a significant loss in precision so that, as the database grew, the search time using the algorithm would not become too long. In 1983, Wilbur and Lipman presented an algorithm called DEC KL-10 that performed such a task. In their original experiment, the authors compared all sequences in the Protein Data Bank, which, at that time, had about 200 entries, with a 350-residue query sequence. The process took about three minutes to complete. A faster algorithm called FASTP was developed in 1985 by Lipman and Pearson, which could do the same operation with about 1000 entries in about 2 minutes. Since these first attempts, the process speed and accuracy have increased significantly with algorithms such as BLAST (Altschul et. al., 1990) and PSI-BLAST (Altschul et. al., 1997) being able to perform this type of search on about 37,500 protein sequences in about 10 to 20 seconds. We will look at BLAST in Chapter 4.

**3.10 Bioinformatics Websites.** The trend of offering publicly available information has been growing steadily in bioinformatics, from books to computer databases of protein sequences, to algorithms able to compare individual sequences to each other as well as to entire databases. The major resource for all databases and tools available to bioinformatics researchers nowadays is the National Center for Biotechnology Information, which was created in 1988. NCBI is a division of the National Library of

Medicine (NLM) at the National Institutes of Health (NIH). NLM was chosen due to its experience in creating and maintaining biomedical databases. Almost any biotechnological database can be found using NCBI, and it also contains a variety of bioinformatics programs. NCBI is currently accessible by anyone with Internet access at <http://www.ncbi.nlm.nih.gov>. When I look at BLAST in chapter 4, the examples will come from using tools within NCBI.

Similar databases have been created since NCBI, including the European Molecular Biology network (EMBL) in 1988, the European Molecular Biology Laboratory in 1994, the DNA Databank of Japan (DDJB) in 1986, and Ensembl in 1999. These databases, along with NCBI, offer a wide range of information and tools. More specialized databases have also been created in order to suit researchers with particular research needs. These include GenBank (Benson et. al., 2005), which acts as a repository for DNA sequences, Protein Information Resource (Wu et. al., 2003), which is an integrated database on protein research, and the Protein Database (Berman et. al., 2000), which stores X-ray crystallographic structures of proteins. There are databases that specialize in particular species, such as the Human Protein Reference Database (Peri et. al., 2003), FlyBase (Ashburner & Drysdale, 1994) and WormBase (Stein et. al., 2001). Lastly, an even newer trend is the creation of databases that specialize in particular molecular functions, such as Reactome (Joshi-Tope et. al., 2005), which describes fundamental molecular pathways in humans, and HumanCyc (Stein, 2005), which describes pathways involved in human metabolism. All of these databases are primarily accessed using the Internet, and each offer a variety of publicly accessible tools for analyzing these databases.



NCBI **Structure**

PubMed Entrez BLAST OMIM Books TaxBrowser Entrez Structure

Search Entrez Structure for  Go

**What's New?**

**MMDB**  
NCBI's structure database

**Cn3D v4.1**  
3D-structure viewer

**CDD**  
Conserved Domain Database

**VAST**  
Structure comparisons

**VAST Search**  
Submit structure database searches

**PDBeast**  
Taxonomy in MMDB


**Research**  
Research topics and staff

Updated 04/22/05

**The NCBI Structure Group**

... maintains MMDB, a database of macromolecular 3D structures, as well as tools for their visualization and comparative analysis. MMDB, the Molecular Modeling Database, contains experimentally determined biopolymer structures obtained from the Protein Data Bank ([PDB](#)).

**Structure highlights**

 A new version of the macromolecular structure viewer Cn3D is now available. New features in Cn3D 4.1 include: new and improved alignment algorithms, VAST alignment import, taxonomy viewer, and a native Mac OSX version. More...

Text searches in MMDB (use the search toolbar at the top of this page) will yield Structure Query pages, providing access to entries that matched the keywords. Structure Summary pages for one/several/all of these may be retrieved. From the Structure Summary pages one may:

- ◆ access amino acid and nucleic acid sequences
- ◆ retrieve PubMed documents
- ◆ get taxonomy information
- ◆ view sequence neighbors
- ◆ view structure neighbors

**try also:**

[Structure summary via PDB/MMDB-Id:](#)  
 Go

Read more about:  
[MMDB](#)  
[WWW-Entrez](#)  
[VAST](#)

Resources:  
[MMDB's FTP-site](#) (including the MMDB database)  
[NCBI C++ Toolkit](#) containing Cn3D source code.  
[Publications in Entrez](#)  
Research software:  
[PKB and Threading](#) (requires Splus).

[Privacy statement](#)  
[Disclaimer](#)

Help Desk NCBI NLM NIH Credits

**Figure 3.5.** Screenshot of main page from NCBI from November 22, 2005.

**3.11 NCBI.** The more general websites, like NCBI, have specific pages that have a variety of databases and tools. Each of these is accessible from the main page (figure 3.5, found on previous page). The main menu bar consists of the following options: Pubmed, Entrez, BLAST, OMIM, Books, TaxBrowser and Entrez Structure. *Pubmed* provides access to biomedical literature. This literature is almost fully integrated into all the tools available on the website. For example, if a researcher wishes to find a protein sequence in the database, journal articles that provide information on the sequence are returned along with the sequence itself. *Entrez* is used to search across all databases that NCBI has access to. For example, if one were to input “Hemoglobin” in the Entrez search field, information on all hemoglobin sequences, DNA, RNA and protein, from every inputted organism, would be returned, along with Pubmed articles on hemoglobin.

Of course, biologists can limit their search in whatever way they wish, such as returning only protein results, or sequences from a specific species, or even results that were input after a particular date. BLAST is a sequence-matching tool already described in this chapter, and we will look into it in more detail in chapter 4. OMIM stands for Online Mendelian Inheritance in Man. It is a catalog of human genes and human genetic disorders. TaxBrowser, short for Taxonomy Browser, provides detailed information on every catalogued species, as well as links to genes and proteins that have been catalogued, and evolutionary relationships to other species (figure 3.6). Finally, Structure is a database specifically designed for storing 3D macromolecular structures, with tools for structure visualization and comparative analysis between structures. Although I will not be going into detail on the prediction of protein structure, it is a highly productive area within bioinformatics. X-ray crystallography is still used to predict protein structure,

but the process is relatively slow, especially when compared to the potential discovery speed that an accurate algorithm can perform. *Ab initio* programs predict protein shape based on the primary structure of proteins and applying rules of folding generated from the folding behaviour of previous proteins with similar amino acids.

**Escherichia coli**

Taxonomy ID: 562  
 Rank: species  
 Genetic code: Translation table 11 (Bacterial and Plant Plastid)  
 Other names:  
 synonym: "Bacterium coli" (Migula 1895) Lehmann and Neumann 1896  
 synonym: Bacterium coli  
 synonym: "Bacillus coli" Migula 1895  
 synonym: Bacillus coli  
 synonym: "Bacterium coli commune" Escherich 1885  
 synonym: Bacterium coli commune  
 synonym: Escherichia coli (Migula 1895) Castellani and Chalmers 1919  
 includes: Escherichia coli retron Ec86  
 includes: Escherichia coli retron Ec79  
 includes: Escherichia coli retron Ec67  
 includes: Escherichia coli retron Ec107

[Lineage \(full\)](#)  
 cellular organisms; Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacteriales; Enterobacteriaceae; Escherichia

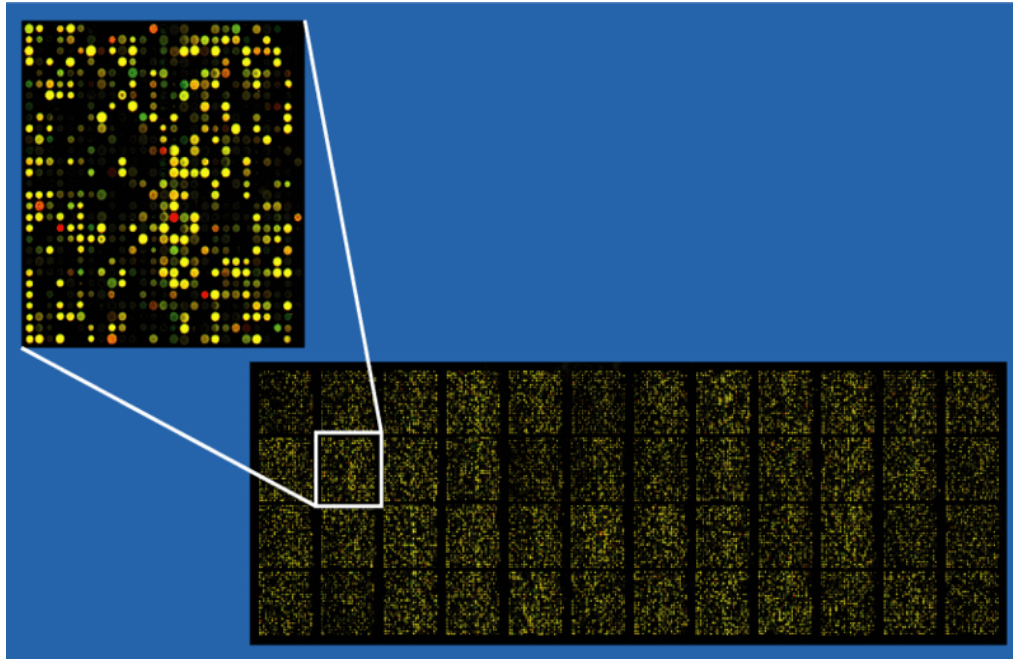
Entrez records		
Database name	Subtree links	Direct links
Nucleotide	33,011	30,279
Protein	127,098	38,045
Structure	3,262	3,246
Genome Sequences	35	20
Genome Projects	14	-
Popset	119	119
3D Domains	17,921	17,878
Domains	8	8
GEO Datasets	65	65
GEO Expressions	50,022	42,710
PubMed Central	68,241	67,032
Gene	21,829	789
Taxonomy	37	1

**Figure 3.6.** *E. coli* information from TaxBrowser. Also contained on this page, but not show here, is a list of references describing the organism in detail and a list of website specialized in the molecular biology of *E. coli*.

**3.12 Microarrays.** One of the most recent and influential developments in bioinformatics is the invention of the microarray, or gene chip, in 1995 (Schena et. al., 1995). The concept of the chip is very simple, yet the technology that is required to facilitate such a development has only been around very recently. The purpose of a microarray is to

monitor the genetic activity of a cell, tissue, or organism under particular conditions and/or a particular time. Previous studies could only monitor the activity of one or a few genes at a time, the most successful technique being the monitoring of what occurs among organisms with mutations of that gene. The problem with this approach, however, is that it was not clear whether those particular genes were the only causes of the effects they observed.

Microarrays circumvent this problem by being able to monitor the activity of all genes at once. They work by placing the same copies of all the genes of a cell, tissue, or organism on each of a number of chips. Then the mRNA products, which are the precursors to proteins, from differently expressed cells/tissues/organisms, are added to each of the chips, and since mRNA are complements of DNA, they bind to the DNA that is found on the chips. For example, one of the chips would have the mRNA from a normal cell added to it and another mRNA from a cell with a particular mutation, or with a drug added, or at a different developmental stage. By applying all the mRNA to chips with all the genes, researchers can monitor all the differences among the types of cells/tissues/organisms. This process is conducive to creating a more accurate picture of the genetic causes and effects within a cell since all of the genes and proteins are being monitored at once. Figure 3.7 shows a set of microarrays. Microarrays will also be studied in greater detail in chapter 6.



**Figure 3.7.** A set of microarrays show the gene expression differences among cells in various different conditions. Each position on one chip corresponds to the same gene on another chip, and any difference in colour corresponds to a difference in the expression of that gene under different conditions. This image is available in the public domain.

**3.13 Summary.** Bioinformatics has three main branches. The first is molecular biology, where researchers provide data on the molecular biology of species. These data include DNA and protein sequencing, testing gene expression with tools like microarrays, and creating better methods for generating such data. The second branch is information theory, which is concerned with the mathematical relationships that exist among genes, substitution frequencies (PAM), and molecular structures. The last branch is computer science, which applies mathematical inferences to the collected biomolecular data, and thus allows scientists to develop biological theories. The computational branch has also taken on another role, and that is making bioinformatics data and tools available to the

wider public, as was shown in this chapter with the numerous websites devoted to bioinformatics.

## Chapter 4

# BLAST Case Study

**4.1 Introduction.** BLAST (Basic Local Alignment Search Tool) (Altschul et. al., 1990) is one of the most extensively used bioinformatics programs. On an average day, the tool is accessed online 200,000 times on the NCBI website (although the tool is found on many other bioinformatics websites, as well), and is said to be used by “every biologist today” (Harding, 2005, p.21). There are two separate reasons for its popularity. The first is that it helps to solve many of the most important current biological problems. These include: finding sequence similarities between input sequences and sequences that are already stored within a database; determining the species origin of sequences that have already been sequenced; and finding genes that share the same domains with an input sequence. Each of these processes will be explained later in this section. The second reason for its popularity has to do with the added scientific reliability and power BLAST has when compared to previous methods. BLAST has both increased the precision of particular biological results and increased the speed of generating these results.

This chapter begins with a history of BLAST and a detailed description of how BLAST works, its underlying algorithms, and the important biological problems it solves. Following the historical and technical descriptions of BLAST, I will show how the use of this tool is an instance of extended cognition as well as analogical reasoning, and show how the tool meets the epistemic standards set by Goldman (1992) and Thagard (1997).

**4.2 What is BLAST?** BLAST is a powerful tool for comparing a protein or DNA sequence to other sequences in various databases. It employs a basic sub-element that is used to compare individual sequences, yet the advantage of BLAST is that it can perform this comparison on potentially millions of sequences, and does it in a relatively short amount of time. The typical BLAST search takes seconds to complete, and can be performed by any individual with Internet access. One simply visits <http://www.ncbi.nlm.nih.gov/BLAST>, and selects one of the BLAST tools available. These tools allow researchers to compare DNA sequences, protein sequences and entire genomes.

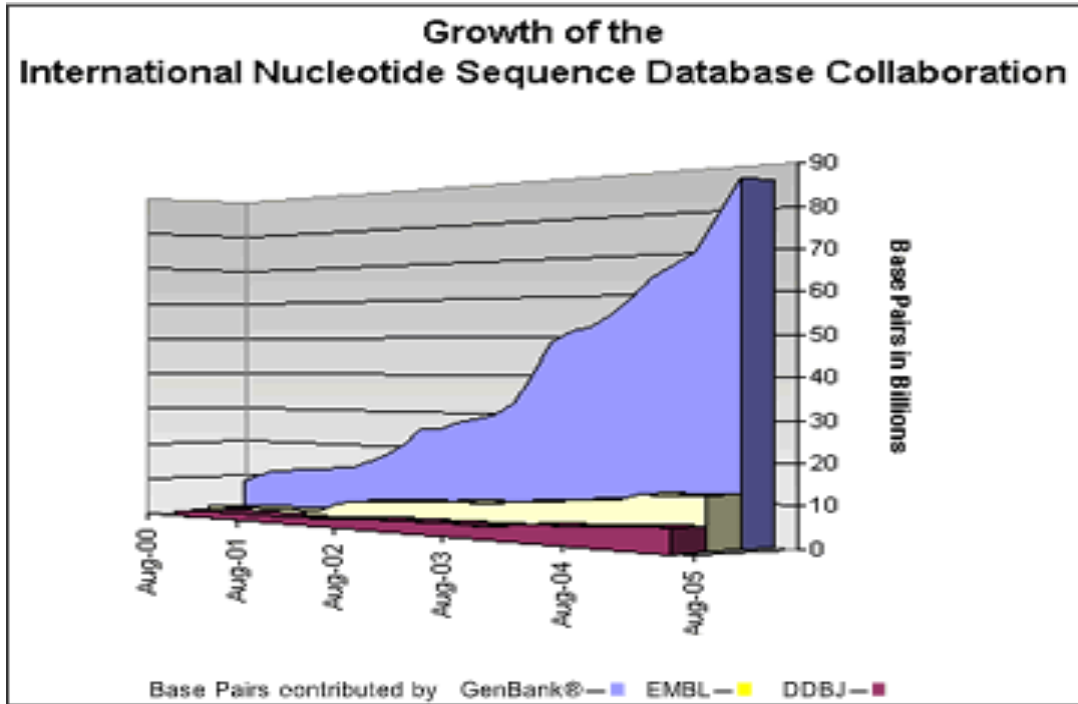
**4.3 Sequence Comparison.** The BLAST algorithm compares individual sequences. The comparison of sequences is a problem fundamental to computer science. One example of this problem is word searches in word processors. But the most popular example of this problem is the web search engine, such as Google.

Any computational method used to compare sequences depends on many different factors: the type of sequences being compared, and what is being sought after the comparison is performed. The first, most intuitive method to compare sequences would be to compare each individual unit of the input sequence with each individual unit of a target sequence. Use this method for the following example: find abcd in khjdafgkabcdjabckdfg. A program using this basic solution would use the first unit in the query sequence and compare it to every unit in the target sequence. So, 'a' would be lined up with 'k', and then 'h' and then 'j', and so on. Once it finds another 'a', the program would check whether the next unit in the query sequence, 'b', matches up with the next



unit in the target sequence. If it does, then it continues to the next unit in the query sequence, and so on. If not, it continues on the target sequences until it finds another 'a' or it reaches the end, at which point it returns a result of 'no matches'.

The problem with this method is that, for comparisons that are more complex, say a query sequence containing hundreds of units being compared to millions of target sequences with similar unit sizes, the process would require a massive number of operations. The number of individual comparisons required for a query sequence of size  $n$  and a target sequence of size  $m$  would be on the order of  $n \times m$ . Using this method for comparing biological sequences would be highly impractical, especially since biological databases are growing at an exponential rate (see figure 4.1). Currently, the DNA database is 100 gigabases in size (100 billion bases), and the average size of a human gene is 27,000 bases long. If one used the simple method described above, the search would perform on the order of  $27,000 \times 100,000,000,000$  operations (2,700,000,000,000,000, or  $2.7 \times 10^{15}$ ). Even if run on a computer that can perform  $10^{12}$  operations per second, which can be done by the world's fastest parallel computing operations, the process would take 1000 seconds, which is about 15 minutes. Although this amount of time is not impractical for a scientist, there are methods that drastically reduce the amount of time required to perform such searches. Also, this search ignores many unique problems that are found in searches through biological sequences, which we will discuss soon.



**Figure 4.1.** Exponential growth of genetic databases (from <http://www.ncbi.nlm.nih.gov/Genbank/>, April 5<sup>th</sup>, 2006, image is in the public domain).

Keep in mind that the number of operations given above does not include the specific additional operations needed for comparing biological sequences. In biological comparisons, one does not only need to find exact matches. Mathematical analyses are made in order to find the degree of relatedness among inexact matching sequences. Some type of *score* is required after a comparison to give a sense of the relative similarity between sequences. Another contingency that is found in biological comparisons is that there often exist *gaps* in comparisons. For example, visually compare the following two sequences:

Sequence 1: ATGATCGTAGACGAGTTCAA

Sequence 2: ATGATCGGAGTTCAA

These two are very similar except that the second one is missing a large section from the first: it is missing “TAGACGAGT”, which is highlighted in the first sequence. This missing section can be caused by many different evolutionary events: either sequence 1 had the missing sequence deleted in a single evolutionary event to produce sequence 2, or sequence 2 had the missing sequence added in a single evolutionary event, or the deletion and/or addition occurred over a number of evolutionary events in either sequence. Both the reason for the gap’s existence as well as the specific gap itself are important for generating accurate alignments and similarity scores when comparing biological sequences.

**4.4 History of BLAST.** The following history of BLAST was compiled by Harding (2005). In 1982, David Lipman and Tim Havell modified the search tools found in UNIX to search for sequence similarities in DNA sequences. Along with John Wilbur, they came up with an algorithm that was able to search through the existing Protein Data Bank of the National Biomedical Research Foundation (NBRF) in less than three minutes, and the Los Alamos Nucleic Acid Data Base in about two minutes.

This early algorithm was already making significant discoveries. Mike Waterfield's lab used it to show the similarity between a viral oncogene and the gene for human platelet-derived growth factor (Waterfield et. al. 1983). Gene Myers was the first to conceive of the BLAST algorithm back in 1988, when he thought that matching sequences using short strings of letters rather than individual letters would create a faster program. David Lipman had a similar idea, but his method used an even broader heuristic tool, so although the algorithms would sometimes miss particular matches, the program

was much faster. Along with Stephen Altschul, Warren Gish, and Webb Miller, Lipman and Myers created BLAST in 1990 (Altschul et. al., 1990). Another key part of the history of BLAST is the statistical tools that were developed for use with BLAST. These will be discussed later in the chapter.

Further refinements of the program have allowed it to run even faster. One of the greatest obstacles to speed was computing power. Once the resources of the Internet were realized as a potential solution to this obstacle, BLAST was soon made available through NCBI as a web tool. Each step, from sequence matching to using strings to the development of a web tool, is described in detail in the following section.

**4.5 Pairwise Sequence Alignment.** I will begin with how pairwise comparisons were made more efficient while not greatly sacrificing accuracy. The comparison may be between two protein sequences, two DNA sequences or two RNA sequences. These sequences can have any degree of relationship, since it is up to the alignment algorithm to determine whether the sequences match exactly or approximates.

The basic bioinformatics pairwise alignments, such as the Needleman-Wunsch or the Smith-Waterman algorithms, begin by putting the sequences in the first row and the first column of an array, and assigning scores in the cells of the rest of the array based on the scores of cells that are nearby. Take the sequences presented earlier in the chapter and place them in the following array (figure 4.2, in this example, the Smith-Waterman local alignment algorithm is described):

		<b>A</b>	<b>T</b>	<b>G</b>	<b>A</b>	<b>T</b>	<b>C</b>	<b>G</b>	<b>T</b>	<b>A</b>	<b>G</b>	<b>A</b>	<b>C</b>	<b>G</b>	<b>A</b>	<b>G</b>	<b>T</b>	<b>T</b>	<b>C</b>	<b>A</b>	<b>A</b>	
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
<b>A</b>	0																					
<b>T</b>	0																					
<b>G</b>	0																					
<b>A</b>	0																					
<b>T</b>	0																					
<b>C</b>	0																					
<b>G</b>	0																					
<b>G</b>	0																					
<b>A</b>	0																					
<b>G</b>	0																					
<b>T</b>	0																					
<b>T</b>	0																					
<b>C</b>	0																					
<b>A</b>	0																					
<b>A</b>	0																					

**Figure 4.2.** The two sequences to be compared are placed in the first row and column of an array.

Next, we assign values to each cell. The values of each cell are determined by picking whatever value is the highest among the following: if a match, then the value of the cell is the value of the cell to its upper left +1; if a mismatch, then, whichever is highest, the value of the cell to its upper left, or left, or above -1; or 0. The process starts with the cell in the upper-left and continues horizontally to the right until the end of the row is reached, and then moves on to the row below until all the cells are filled. The first row would look like (figure 4.3):

		<b>A</b>	<b>T</b>	<b>G</b>	<b>A</b>	<b>T</b>	<b>C</b>	<b>G</b>	<b>T</b>	<b>A</b>	<b>G</b>	<b>A</b>	<b>C</b>	<b>G</b>	<b>A</b>	<b>G</b>	<b>T</b>	<b>T</b>	<b>C</b>	<b>A</b>	<b>A</b>	
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
<b>A</b>	0	1	0	0	1	0	0	0	0	1	0	1	0	0	1	0	0	0	0	0	1	1

**Figure 4.3.** Values of the first unit of sequence 2 when matched to sequence 1.

Notice in each of these cells there is only a match (+1) or a 0. The next row would be the following (figure 4.4):

		A	T	G	A	T	C	G	T	A	G	A	C	G	A	G	T	T	C	A	A
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A	0	1	0	0	1	0	0	0	0	1	0	1	0	0	1	0	0	0	0	1	1
T	0	0	2	1	0	2	1	0	1	0	0	0	0	0	0	0	1	1	0	0	0

**Figure 4.4.** Values of the second row comparing the second unit of sequence 2 with each unit of sequence 1. Some of the values are dependant upon the values arrives at in the previous row.

Notice that some of the cells have '2' since the match in those cells is added to the match in the cell upper-left of it. The cells to the right of the cells with '2' have a '1' since they get a -1 penalty from the cell to its left. Continuing this strategy, we get (figure 4.5):

		A	T	G	A	T	C	G	T	A	G	A	C	G	A	G	T	T	C	A	A
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A	0	<b>1</b>	0	0	1	0	0	0	0	1	0	1	0	0	1	0	0	0	0	1	1
T	0	0	<b>2</b>	1	0	2	1	0	1	0	0	0	0	0	0	0	1	1	0	0	0
G	0	0	1	<b>3</b>	2	1	0	2	1	0	1	0	0	1	0	1	0	0	0	0	0
A	0	1	0	2	<b>4</b>	3	2	1	1	2	1	2	1	0	2	1	0	0	0	1	1
T	0	0	1	1	3	<b>5</b>	4	3	2	1	1	1	0	0	1	1	2	1	0	0	0
C	0	0	0	0	2	4	<b>6</b>	5	4	3	2	1	2	1	0	0	1	0	2	1	0
G	0	0	0	1	1	3	5	<b>7</b>	6	5	4	3	2	3	2	1	0	0	1	0	0
G	0	0	0	1	0	2	4	6	<b>6</b>	5	6	5	4	3	2	3	2	1	0	0	0
A	0	1	0	0	2	1	3	5	5	<b>7</b>	6	7	6	5	4	3	2	1	0	1	1
G	0	0	0	1	1	1	2	4	4	6	<b>8</b>	<b>7</b>	<b>6</b>	<b>7</b>	<b>6</b>	<b>5</b>	4	3	2	1	0
T	0	0	1	0	0	2	1	3	5	5	7	6	6	5	<b>6</b>	<b>5</b>	<b>6</b>	5	4	3	2
T	0	0	1	0	0	1	1	2	4	4	6	5	5	4	5	4	6	<b>7</b>	6	5	4
C	0	0	0	0	0	0	2	1	3	3	5	4	6	5	4	4	5	6	<b>8</b>	7	6
A	0	1	0	0	1	0	1	1	2	4	3	6	5	5	6	5	4	5	7	<b>9</b>	8
A	0	1	0	0	1	0	0	0	1	3	5	5	5	4	6	5	4	4	6	8	<b>10</b>

**Figure 4.5.** The completed array. The bolded numbers represent the optimal alignment of the two sequences based on this algorithm.

The numbers that are bolded indicate the optimal alignment of the two sequences as determined by the algorithm. This is done by starting at the cell with the highest score (in this case, it is at the bottom right with a score of 10) and selecting the cell above, to the left or to the upper-left, whichever has the higher score. This continues until a '0' is reached. When the order of bolded cells is diagonal, then it means that the units in each sequence are aligned without any gaps. When the order is vertical or horizontal, then that means that there is a gap in one of the sequences: a horizontal order means a gap in the vertically positioned sequence on the array, and a vertical order means a gap in the horizontally positioned sequence on the array:

```
Alignment #1:      Sequence 1: ATGATCGTAGACGAGTTCAA
                   | | | | | | | | | | | | | | | | | | | | | |
                   Sequence 2: ATGATCGGAG - - - - - TTCAA

Alignment #2:      Sequence 1: ATGATCGTAGACGAGTTCAA
                   | | | | | | | | | | | | | | | | | | | | | |
                   Sequence 2: ATGATCGGAG - - - T - - TCAA
```

The vertical bars indicate unit identity between the two sequences. Notice there is at least one position where the units are not identical, and there is at least one gap in each of the two optimally calculated alignments.

**4.6 Biological Sequence Alignments.** With the method of placing the sequences in arrays, scoring them and finally finding alignments based on those scores, we have achieved a necessary step in accounting for the contingencies found in biological sequence alignments, such as the inexact matchings that are found due to single

nucleotide polymorphisms and gaps. Looking at the alignments above, we see a problem: there are two optimally calculated alignments predicted by the algorithm instead of only one. Which one would be *the* optimal alignment?

Optimal alignments can be defined in at least two ways. Mathematically, the optimal alignment is found “by making a series of decisions at each step of the alignment as to which pair of residues corresponds to the best score.” (Pevsner, 2003, p.67). Biologically, however, the optimal alignment is that which matches the actual biological similarities between the sequences and that accounts for the actual evolutionary changes that have occurred among the sequences. In order to find an optimal alignment in biological sequences, one needs to take these two definitions into account. In this case, the problem can be partially solved when we consider prior knowledge from genetics research: this research tells us that insertions and deletions in DNA sequences, or *indel mutations*, are rather rare events. Indel mutations cause large changes in the protein that is created, therefore potentially causing large phenotypic changes. Thus, when choosing between two calculated alignments, it is more likely that an optimal alignment will have fewer gaps.

Two optimal alignments were calculated above due to the problems in the simple scoring method employed; therefore, a more complex scoring method is needed in order to account for gap rarity and produce only one optimal alignment. For example, gaps should be given a larger penalty than the  $-1$  score we have above, since gaps are more biologically rare than single nucleotide substitutions. Alignment programs even account for a difference between a *gap origin* and *gap extension*. Even though gaps are given large penalties, evolutionarily speaking, that penalty should not be as severe as the gap



extends. In other words, if a phenotypic change would occur from any indel mutation, then the size of that indel should not make a difference in the fact that a phenotypic change would occur anyway. Thus, programs such as BLAST give large penalties for a gap's origin (say -13) but reduce that penalty for each nucleotide that the gap extends (say -6). If this new method had been implemented in the alignment above, only the first alignment would have been optimal.

Another refinement to the algorithm comes from realizing that nucleotides come in two different forms: purines and pyrimidines. Purines are the nucleotides adenine (A) and guanine (G), and pyrimidines are the nucleotides cytosine (C) and thymine (T). When one purine is substituted for another purine, or a pyrimidine is substituted for another pyrimidine, it is called a *transition*, which is more likely than *transversions*, which is when a purine is substituted for a pyrimidine, or vice-versa. Therefore, a scoring method may take this into account, giving greater mismatch scores to transversions (say -2) than to transitions (say -1).

The alignment of nucleotides is informationally important, but protein alignments are much more interesting. This is largely due to the fact that proteins are the molecules that are involved in the actual biological mechanisms of organisms. Another reason is that the genetic code is *redundant*, meaning that various nucleotide sequences encode for the same amino acid (which are the building blocks of proteins). Below (figure 4.6) is the table of the *Universal Genetic Code*, which applies to almost all species. It demonstrates which triplet of the nucleotides A, G, C and U code for which proteins. The three letter names of each protein are in bold, and the one letter version is in brackets. Although this

list is called ‘universal’, it is not, as there have been other amino acids discovered that are species specific, such as hydroxyproline and taurine.

<b>Ala</b> (A) GCU, GCC, GCA, GCG	<b>Leu</b> (L) UUA, UUG, CUU, CUC, CUA, CUG
<b>Arg</b> (R) CGU, CGC, CGA, CGG, AGA, AGG	<b>Lys</b> (K) AAA, AAG
<b>Asn</b> (N) AAU, AAC	<b>Met</b> (M) AUG
<b>Asp</b> (D) GAU, GAC	<b>Phe</b> (F) UUU, UUC
<b>Cys</b> (C) UGU, UGC	<b>Pro</b> (P) CCU, CCC, CCA, CCG
<b>Gln</b> (Q) CAA, CAG	<b>Ser</b> (S) UCU, UCC, UCA, UCG, AGU, AGC
<b>Glu</b> (E) GAA, GAG	<b>Thr</b> (T) ACU, ACC, ACA, ACG
<b>Gly</b> (G) GGU, GGC, GGA, GGG	<b>Trp</b> (W) UGG
<b>His</b> (H) CAU, CAC	<b>Tyr</b> (Y) UAU, UAC
<b>Ile</b> (I) AUU, AUC, AUA	<b>Val</b> (V) GUU, GUC, GUA, GUG
<b>Start</b> AUG, GUG	<b>Stop</b> UAG, UGA, UAA

**Figure 4.6.** The Universal Genetic Code. The amino acid is listed on the left side of each cell. Its corresponding one-letter identification is in brackets.

What this table shows is that some amino acids can be coded by a few different *triplets* of nucleotides. Thus, even if some nucleotide sequences do not exactly align, the sequence may still be phenotypically equivalent, since the same amino acids are translated from those nucleotide sequences. Therefore, looking at protein sequences can give much more information about the relatedness between two sequences, especially if they are more distantly related.

Amino acids also have the interesting characteristic that many of them have similar properties to each other. The twenty amino acids are grouped into four major groups: non-polar (hydrophobic), uncharged polar, negatively charged and positively charged. If one non-polar amino acid is substituted for another non-polar amino acid, then the properties and functions of the protein may not significantly change. If, on the other hand, a polar protein is substituted for a non-polar one, then the protein’s properties



amino acid. This effect is also observed in the significantly high mismatch scores for tryptophan when compared to the significantly low mismatch scores for serine.

These changes, which include complex gap penalties and specific scoring matrices, are the most important technical developments for algorithms that perform biological sequence alignments in attempting to find optimal alignments. However, these developments do not completely remove one of the problems we encountered at the beginning of this chapter: that of speed. In the classic Smith-Waterman and Needleman-Wunsch algorithms, every nucleotide is still being compared to every other nucleotide, which means that the process requires that there be at least  $n \times m$  operations. This is a lot when a sequence is being compared to the entire genetic database. BLAST makes this process quicker, but understanding how BLAST works is not possible without knowing the basic processes described so far.

**4.7 The BLAST Algorithm.** This section will describe the blastp algorithm, which is the BLAST algorithm for protein alignments. Blastp takes advantage of the different matching scores that are specific to biological sequencing, such as the scores found on the PAM array. The first step in the blastp algorithm is to split the query sequence into ‘words’. The typical word size used in blastp is 3. For example, take the following protein sequence, taken from Pevsner (2003, p.101):

Human Retinol Binding Protein: ...FSGTWYAMAKKDP...

The portion of the sequence shown here is then divided into overlapping words of size 3:

FSG SGT GTW TWY WYA YAM AMA MAK AKK KKD KDP

For each of these words, similar words are found for which the match score is above some *threshold* value. Typically, BLAST programs use a threshold of +11 for protein sequences. Words that have a value below the threshold are discarded, and those equal and above are kept. For example, if the word is GTW, the set of words above the threshold is:

$$\text{GTW } (6 + 5 + 11 = 22)$$

$$\text{GSW } (6 + 1 + 11 = 18)$$

$$\text{GNW } (6 + 0 + 11 = 17)$$

$$\text{GAW } (6 + 0 + 11 = 17)$$

$$\text{ATW } (0 + 5 + 11 = 16)$$

$$\text{DTW } (-1 + 5 + 11 = 14)$$

$$\text{GTF } (6 + 5 + 1 = 12)$$

Threshold (11) -----

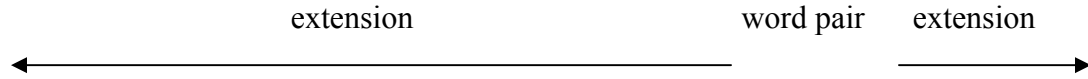
$$\text{GTM } (6 + 5 - 1 = 10)$$

$$\text{DAW } (-1 + 0 + 11 = 10)$$

...

Note: the score for each word is pre-computed, as these are 8,000 protein words of size 3. Once the words that are above the threshold value are found, one searches the protein or DNA databases to find sequences with the same words, which serve as “seeds”. The matched sequences are then scanned from the seed word in both directions of the matching sequence using the pairwise alignment method described in a previous section, the Smith-Waterman algorithm. The scan stops when the score drops past a certain cut-off value or when the end of either sequence is reached:

MKWVWALLLLAAWAAAERDCRVSSFRVKENFDKARFSG**TWY**AMAKKDPEG  
 .....MKCLLLALA L TCGAQAL IVT...QTMKGLDIQKVAG**TWYS** LAMAASD..



The bolded letters comprise the original word match, and then the pairwise alignment proceeds along in both directions. The lines between the sequences indicate an identical match between the two residues. The dots indicate similar matches, i.e., amino acids that have similar properties and correspond to a positive score in the scoring matrix.

Once all alignments are found, the significance is estimated using the scores from the pairwise alignments. I will not go into how the probabilities are calculated, but the statistics for BLAST do not employ the normal distribution curve, but an *extreme value* curve instead. This is due to the fact that if a query sequence is compared to a set of random sequences of equal length, the scores that are generated have an extreme value distribution. Thus, instead of the normally used p-values found in statistical analyses, BLAST uses *Expect values* or *E-values* for short. An E-value corresponds to the number of different alignments with scores the same or better than the score, *S*, that is expected to occur by chance in a search in a protein or DNA database.

The BLAST method reduces the time of sequence searches dramatically. Although two sequences are still being matched letter by letter, one of the sequences has only 3 letters in the case of proteins and 11 letters in the case of DNA. For DNA, this reduces the number of operations from  $2.7 \times 10^{15}$  to  $1.1 \times 10^{11}$ , multiplied by the number of words that are above the threshold value. Once matching sequences are found, the

sequences are compared only until the score drops below some value or when the end of a sequence is reached. Current popular BLAST programs have incorporated many other strategies to further reduce search time, but I will not cover them here, since the main strategy of BLAST is all that is needed to understand the following sections. Thus, the main trade-off that is made using the BLAST algorithm is that of speed for accuracy. However, the loss of accuracy can be seen as acceptable when one considers the gain in speed that is made.

**4.8 The BLAST Search.** I will now describe how BLAST searches are performed using various specifically designed websites. The purpose of presenting these websites is to show the ease and power of using these computational tools, which will figure importantly once we get to the methodological evaluation.

The most popular bioinformatics website in North America is the “National Center for Biotechnology Information” at <http://www.ncbi.nlm.nih.gov>. Once one visits the website, one can select ‘BLAST’ from the top menu bar. On the BLAST page, there are many different options one can select. There are six main options: searches can be conducted 1) using nucleotide sequences, 2) using protein sequences, 3) using translated sequences (DNA translated to protein), 4) by looking through particular genomes (say human or mouse), 5) using special tools that are not categorizable in the previous 4 categories, and 6) using a meta search. The nucleotide and protein searches can also be combined. The following are frequent types of BLAST searches:

blastp: protein sequence inputted and matching protein sequences returned.

blastn: DNA sequence inputted and matching DNA sequences returned.

blastx: DNA sequence inputted and matching protein sequences returned.

tblastn: protein sequence inputted and DNA sequences returned.

tblastx: DNA sequence is inputted and translated into its 6 possible proteins (since DNA has 6 possible reading frames), and then those proteins are matched to DNA sequences, which are returned.

A blastp query will now be described. On the next page is the blastp input page (figure 4.8). From here, one inputs the protein sequence and modifies the parameters in order to get one's desired output. The following page can be found at <http://www.ncbi.nlm.nih.gov/BLAST/> and by selecting 'Protein-protein BLAST (blastp)'. I have chosen to show the blastp page since it is among the most popular of the BLAST algorithms because:

- Protein sequences give better indications of evolutionary relationships due to the redundancy of the genetic code,
- The sequences allow for the comparison of sequences that are more distantly related, and
- The sequences allow for quicker search times since they are shorter than DNA and RNA counterparts.



[Search](#)

[Set subsequence](#) From:  To:

[Choose database](#)

[Do CD-Search](#)

Now: **BLAST!** or [Reset query](#) [Reset all](#)

**Options** for advanced blasting

[Limit by entrez query](#)  or select from:

[Compositional adjustments](#)

[Choose filter](#)  Low complexity  Mask for lookup table only  Mask lower case

[Expect](#)

[Word Size](#)

[Matrix](#)  Gap Costs

[PSSM](#)

[Other advanced](#)

[PHI pattern](#)

**Format**

Show  Graphical Overview  Linkout  Sequence Retrieval  NCBI-gi Alignment  in  format

CDS feature

[Masking Character](#)  [Masking Color](#)

Number of: [Descriptions](#)  [Alignments](#)

[Alignment view](#)

[Format for PSI-BLAST](#)  with inclusion threshold:

[Limit results by entrez query](#)  or select from:

[Expect value range:](#)

[Layout:](#)  [Formatting options on page with results:](#)

[Autoformat](#)

**BLAST!** or [Reset all](#)

Get the URL with preset values ? [Get URL](#)











**Figure 4.8.** The Blastp input screen from Feb. 12<sup>th</sup>, 2006.

This form allows the user to specify how the search should proceed.. The ‘Search’ field is where the protein sequence is input. The sequence that is input normally follows the ‘FASTA’ format, which has the one-letter designation for each amino acid. For example, Methionine is designated M and Lysine is designated K. The full list of designations can be found in the Universal Genetic Code figure presented earlier in this chapter (figure 4.6). A researcher can perform a BLAST search as soon as this field has an input sequence, but a researcher may wish to change some of the default parameters. The ‘Choose database’ parameter allows researchers to select which database to search. The default database is the ‘non-redundant’ database, but a researcher may choose ‘PDB’ in order to only have proteins with known 3-D structures returned. The search can also be limited to a particular kingdom of life, or even to a particular species. For example, a researcher who wishes to discover whether a newly sequenced mouse gene has an analogue in humans can restrict the search to ‘Homo sapiens’. The cut-off Expect value can be changed in order to return more or less results from a query. The default cut-off E-value is 10, but if it were changed to say, 20, then the query will return more results. This is a useful strategy in case one gets too many results from a query and wishes to refine the search, or too few results from a query and wishes to find more matches. The word size, match-score array (like PAM or BLOSUM), gap penalty and gap extension penalties can also be changed in order to modify the results one receives. These are only a few of the changes one can make, and combined with the others, the blastp web page has proven to be very flexible and can meet the research demands of most bioinformaticists.

If the mutated human coagulation factor VIII protein sequence, a sequence which causes for hemophilia, is input into the blastp search space along with the default search

parameters, the results below (figure 4.9) are returned. These results are edited since the actual results would take about 20 thesis pages.

### Related Structures

Sequences producing significant alignments:	Score (Bits)	E Value	
<a href="#">gi 4503647 ref NP_000123.1 </a> coagulation factor...	<a href="#">4657</a>	0.0	
<a href="#">gi 31499 emb CAA25619.1 </a> unnamed protein product...	<a href="#">4654</a>	0.0	
<a href="#">gi 182383 gb AAA52420.1 </a> coagulation factor VIII	<a href="#">4653</a>	0.0	
<a href="#">gi 182803 gb AAA52484.1 </a> factor VIII	<a href="#">4652</a>	0.0	
<a href="#">gi 224258 prf 1012298A</a> factor VIIIC	<a href="#">4651</a>	0.0	
<a href="#">gi 66773789 sp O18806 FA8_CANFA</a> Coagulation fact...	<a href="#">3560</a>	0.0	
....			
<a href="#">gi 27806943 ref NP_776304.1 </a> coagulation factor V	<a href="#">525</a>	9e-147	
<a href="#">gi 163040 gb AAA30513.1 </a> factor V	<a href="#">525</a>	9e-147	
<a href="#">gi 6679731 ref NP_032002.1 </a> coagulation factor V	<a href="#">524</a>	2e-146	
<a href="#">gi 50513523 pdb 1SDD B</a> Chain B, Crystal Structure	<a href="#">524</a>	2e-146	
<a href="#">gi 16200178 emb CAC94896.1 </a> novel protein similar	<a href="#">523</a>	4e-146	
<a href="#">gi 55588728 ref XP_513984.1 </a> PREDICTED: coagulati	<a href="#">498</a>	9e-139	
....			
Score = 211 bits (538), Expect = 2e-52			
Identities=119/350 (34%), Positives=178/350 (50%), Gaps=18/350 (5%)			
1749 FKKVVFQEFDTGSPFTQPLRYGELNEHLGLLGPYIRAIVEDNIMVTFRNQASRPYSFYSSL			
1808 +KK V++++TD ++T + + LG LGP IRAEV D I V +N ASRPY+ +			
40 YKKSVMYKQYTDSTYTTTEIPKPAW---LGFLGPIIRAIEVGDITIKVHLKNFASRPYTIHPHG 96			
1809 ISYEEDQRQGAEP-----RKNFVKPNETKTYFWKVQHMMAPTKEFDCKAWAYFSDV			
1860 + YE+ P + + V P + TY W V +PT D+ +C W Y S +			
97 VFYEKSGSEGLYPDMSPQDQKDDAVFPGGSYTYTWTVPEDHSPTADDPNCLTWIYHSHI 156			
1861 DLEKDVHSLIGPLLVCHTNTLNPAGRQVTVQ-EFALFFTIFDETKSWYFTENMERNCR			
1919 D KD+ SGLIGPL+ C L R+ V +F L F++ DE SWY EN+ C			
157 DAPKDIASGLIGPLVTCKEGILTGTQRQDQDQVDFFLMFSVVDENLSWYLDENIASFCT 216			
1920 APCNIQMEDPTFKENYRFHAINGYIMDTLPGLVMAQDQIRIRWYLLSMGSNENIHSIHFSG			
1979 P ++ ED F+E+ + HAING++ LP L M + W+L MG+ +IH+ +F G			
217 DPGSVKDEDEFQESNMKHAINGFVFGNLPALTMACAGDHVAWHLFGMGNEIDIHTAYFHG 276			
1980 HVFTVRKKEEYKMALYNLYPGVFETVEMLPSKAGIWRVECLIGEHLHAGMSTLFLVYSNK			
2039 ++R ++ + +L+P F T +M+P G W + C + +H+ AGM+ ++ V			
277 ETLSIR---GHRTDVASLFPATFVTADMIPGNPGRWLLSCLNDHIQAGMAAIYEVRPCS 333			

**Figure 4.9.** Results from BLASTP search comparing human coagulation factor VIII with the rest of the proteins stored in the database.

The first section lists the proteins that are similar to the input sequence and made the E-value cutoff of 10. I presented a few that had an E-value of 0.0, meaning that these sequences are almost identical to the query sequence, as well as others that had a slightly larger E-value, meaning that there are some dissimilarities between the two sequences. In all, 1478 alignments were returned for this query. Included in this list, from left to right, is the gene identification number, the protein identification number, the function of the protein, the score, the E-value, whether a corresponding gene has been sequenced (represented by the boxed 'G') and whether there is a predicted structure for the protein (represented by the boxed 'S').

After listing the similar sequences, the returned results show the pairwise sequence alignments of the query sequence with all the returned results. The alignments are similar to the one presented earlier in this chapter. The percent identity, similarity and gaps are also calculated. The numbers beside each row of amino acids indicate the amino acid position in the protein. These returned results allow the researcher to look at each alignment and make some judgments as to whether the alignment is acceptable or not, i.e., whether some of the parameters on the input page need to be changed. Due to the massive number of returned results, the researchers may also wish to make the query conditions stricter. This summary provides a basic understanding of how BLAST works. Next, we will look at particular cases of the use of BLAST in scientific research.

**4.9 The BLAST Case.** One area in which BLAST has been significantly helpful is in initiating new avenues for human medical research. Specifically, BLAST has been used to find genes in other species that are homologous to human genes, particularly those that

are implicated in various diseases. Animal models have been used by medical researchers in testing new treatments, and BLAST can help refine those models by finding animals that have similar biological entities and mechanisms that are under investigation. Rubin et. al. (2000) performed a major genetic comparative study among the fully sequenced eukaryotic genomes, and found proteins homologous to human disease proteins in the fruit fly (*Drosophila melanogaster*), the flatworm (*Caenorhabditic elegans*) and in yeast (*Saccharomyces cerevisiae*). With BLAST, researchers can perform experiments on the homologous disease proteins on the model animals in order to discover novel treatments.

Rubin et. al. performed their disease protein comparisons using blastp, which is the BLAST tool used to compare protein sequences. Protein comparisons are the most informative in this case for several reasons: 1) Due to the historical divergence between the compared species, it is likely that many nucleotide substitutions occurred, but that these changes did not result in relevant protein changes because of the genetic code being redundant (as we saw in a previous section). 2) Other genetic changes, such as the insertion of introns<sup>1</sup>, can cause genotypic differences that are not translated in the organism's proteins. 3) Gene annotation is still very inaccurate, meaning that researchers are much more confident about the function of particular proteins than they are about particular genes.

The researchers compiled a list of 289 human proteins that have particular mutations, alterations, amplifications, or deletions that have been implicated in various human diseases. Of these 289 genes, they found that 177 (61%) have homologs in the

---

<sup>1</sup> Introns are portions of a gene that are not normally translated into proteins. The portions that are translated are termed 'exons'.

fruit fly, where homology meant that there was E-value equal to or less than  $10^{-10}$  for 80% of the alignment. Some proteins that were not found to have homologs in the fruit fly were absent because these proteins play roles in biological systems that are absent in the fruit fly. For example, hemoglobin, when mutated, can cause hemophilia and sickle cell anemia. However, this protein is absent in the fruit fly since insects do not require oxygen-transport erythrocytes. In these species, oxygen is delivered directly to their cells via the tracheal system, which is simply a system of open-air tubes.

Of the genes that were found to be homologous, studying the pathways of these genes in the fruit fly may be useful in the study of their human counterparts. For example, of the cancer genes surveyed, 68% had homologs in the fruit fly. Included is p53, which is the gene that, when mutated, is part of the cause of many cancer cases. Since most forms of cancer affect cells regardless of tissue type, studying the roles of p53 in the fruit fly may be helpful for studying cancer in humans, even though many of our organ systems are very different. Other cancer genes that were found to be homologous include menin (MEN), Peutz-Jeghers disease (serine-threonine kinase 11, or STK11), ataxia telangiectasia (ATM) and multiple exostosis type 2 (EXT2).

Rubin et. al. do not present the actual homology searches that were performed since such a presentation would have taken many volumes. In order to get a sense of the searches, I have performed one using comparing the STK11 protein (the protein responsible for Peutz-Jegher's disease) in humans with the fruit fly genome. This is done in order to show how each comparison can be done. Below is the FASTA representation of the STK11 protein, the FASTA format being how genes are represented in bioinformatics databases (also see chapter 3 for a reminder on FASTA formats):

gi|4507271|ref|NP\_000446.1| serine/threonine protein kinase 11 [Homo sapiens]  
 MEVVDPPQQLGMFTEGELMSVGMDFIHRIDSTEVIYQPRRKRAKLIQKYLMDLLGEGSYGKVKVLDSET  
 LCRRAVKILKKKKLRRIPNGEANVKKEIQLLRRLRHKNVIQLVDVLYNEEKQKMYMVMEYCVCGMQEMLDS  
 VPEKRFPVCQAHGYFCQLIDGLEYLHLSQGIHVHKDIKPGNLLLLTTGGTLKISDLGVAEALHPFAADDCRTS  
 QGSPAFQPPEIANGLDTFSGFKVDIWSAGVTLYNITTGLYPFEGDNIYKLFENIGKGSYAIPGDCGPPLSD  
 LLKGMLEYEPAKRFSIRQIRQHSWFRKKHPPAEAPVPIPPSPDTKDRWRSMTVVPYLEDLHGAEDEDEDLFD  
 IEDDIYTDFTVPGQVPEEEASHNGQRRGLPKAVCMNGTEAAQLSTKSRAEGRAPNPARKACSASSKIRR  
 LSACKQQ

This sequence was inputted into the blastp window, and the search was restricted to *D. melanogaster*. The following sequence was the top hit, and the pairwise alignment is shown below the sequence:

[gi|24646654|ref|NP\\_731846.1|](#)lklbl CG9374-PI, isoform I Score= [384](#) E=9e-107

Identities = 204/326 (62%), Positives = 258/326 (79%), Gaps = 5/326 (1%)

Query	23	DTFIHRIDSTEVIYQPRRKRAKLIQKYLMDLLGEGSYGKVKVLDSETLCRRAV	kilkk	82
		+ F +R+DS ++IYQ ++K K++GKY+MGD+LGEYSYKVKVE ++SE LCR AVKIL K		
Sbjct	146	NMFFNRVDSQDIYQKKKSIKMGVGYIMGDVLEGEYSYKVKVEAMNSENLCRLAVKILTK		205
Query	83	kklrrripngeanvkkeiqllrllrhknviqlvdvlyneekqkmyvmeycvcgmqe	mls	142
		+KLRRIPNGE NV +EI LL++L+H++V++LVDVLYNEEKQKMY+VMEYCV G+QEM+D		
Sbjct	206	RKLRRIPNGEQNVTRIEIALLKQLKHRHVVELVDVLYNEEKQKMYLVMEYCVGGLQEMIDY		265
Query	143	VPEKRFPVCQAHGYFCQLIDGLEYLHLSQGIHVHKDIKPGNLLLLTTGGTLKISDLGVAEALH		202
		P+KR P+ QAHGYF QL+DGLEYLHS ++HKDIKPGNLLL+ TLKISD GVAE L		
Sbjct	266	QPDKRMPFLFQAHGYFKQLVDGLEYLHSCRVIHKDIKPGNLLLLSLDQTLKISDFGVAEQLD		325
Query	203	PFAADDCRTSQQGSPAFQPPEIANGLDTFSGFKVDIWSAGVTLYNITTGLYPFEGDNIYK		262
		FA DDTCT QGSPAFQPPEIANG +TF+GFKVDIWS+GVTLYN+ TG YPFEGDNIY+		
Sbjct	326	LFAPDDTCTTGQSPAFQPPEIANGHETFAGFKVDIWSGVTLYNLATGQYPFEGDNIYR		385
Query	263	LFENIGKGSYAIPG---DCGPPLSDLLKGMLEYEPAKRFSIRQIRQHSWFRKKHppaeap		319
		L ENIG+G + P + ++L+ GML+ +P+KR S+++IR +WFR P		
Sbjct	386	LEENIGRGQWEAPAWLYEMDADFANLILGMLQADPSKRLSLQEIIRHDTWFRSAPVKTGPP		445
Query	320	vpippspDTKDRWRSMTVVPYLEDLH		345
		+PIPP D++R+ TV+PYLE H		
Sbjct	446	IPIPPLKG--DKYRNSTVIPIYLEAYH		469

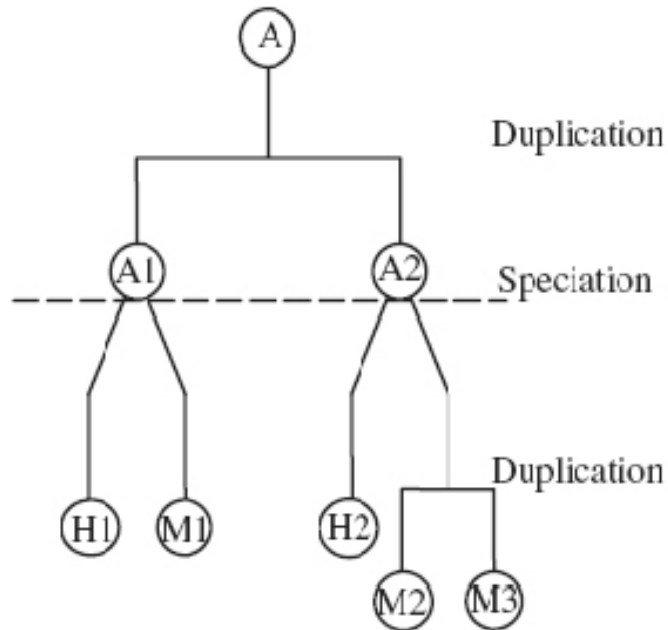
This alignment is especially good, since there is a 62% identity and 79% similarity between the two sequences. These percentages are high, considering the evolutionary divergence between humans and fruit flies. The fruit fly gene would therefore be seen as a good candidate for medical research.

Along with finding information that could be useful for medical purposes, Rubin et. al. also used their resulting homology database to make generalized comparisons of the four eukaryotic genomes. For example, they found that, although the proteome sizes of the flies and worms are only twice that of yeast, the proteins of the former two species are much more complex. Also, the fly and worm proteins are used for a variety of different purposes, both within the cells and in the extra-cellular region.

Since Rubin et. al.'s study, bioinformatics studies have become much more complex. One of the problems with their study is that it does not distinguish between *orthologs* and *paralogs*. Orthologs are what are classically considered homologs, that is, genes that are related through some ancestral species. Paralogs, on the other hand are genes that are related through a gene duplication event within a species, meaning that at some point, a species had duplicate copies of a gene in their genome. These duplicate genes evolve separately, however, and produce genes within the species that may have different functions. For example, hemoglobin is made from a number of genes, each a gene in the 'globin' family. It has been determined that these globin genes arose from a duplication of a single globin gene and each evolved separately to make up the parts of hemoglobin. There are dozens of globin genes in humans; these are paralogs genes, or more specifically *inparalogs*. The globin genes have orthologs counterparts in other mammalian species, for example, globin A in humans has a globin A ortholog in mice. However, only the globin A gene in one species is orthologous to the globin A gene in another species. Globin A in species 1 is said to be *outparalogous* to globin B in another species. This presents a problem in bioinformatics study, since, until these *gene families*

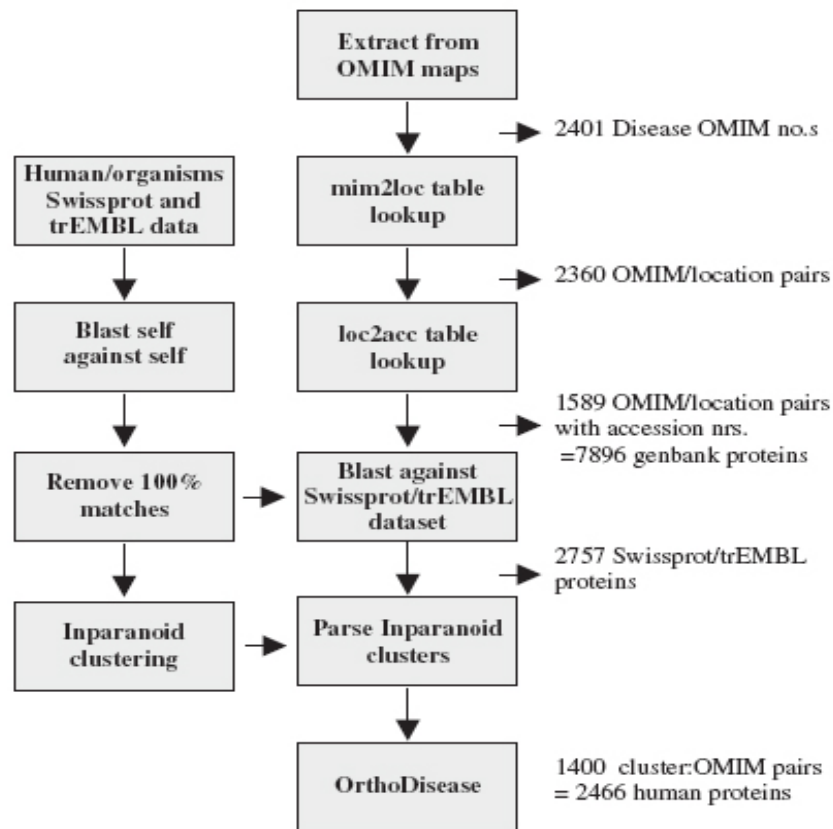


are discovered, it is not clearly known which genes are orthologs and which are merely outparalogs. The following figure presents the various possibilities:



**Figure 4.10.** A hypothetical gene tree to illustrate the relationships leading to inparalog (co-ortholog) and outparalog assignments (from O'Brien et. al., 2004).

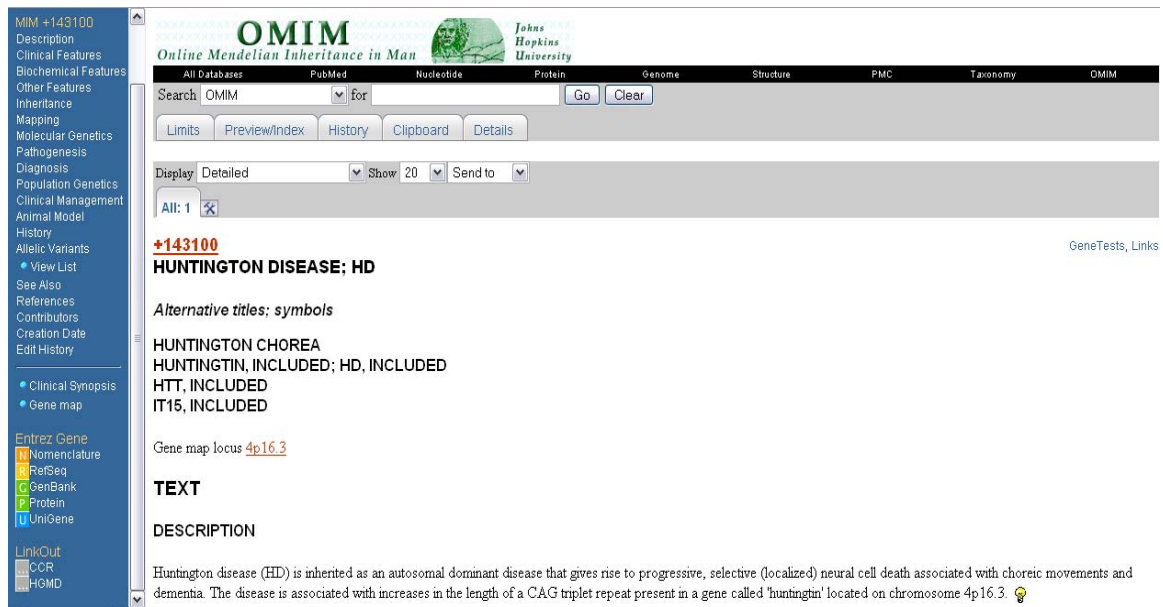
Due to this complexity, many authors have attempted to take the orthologs/paralogs distinction into account. O'Brien et. al. (2004) created 'OrthoDisease', which is designed to only find the orthologous genes that are found between species, since outparalogs may have completely different functions and pathways. This is no easy task, however. Although BLAST is still used to create this database, a variety of other programs are used as well, as we see in the following diagram from the authors (figure 4.11):



**Figure 4.11.** A schema of the algorithms and databases used in order to generate the OrthoDisease database. These extra steps were used in order to distinguish between inparalogs, outparalogs and orthologs, and make sure that the final database only contained orthologs.

One of the databases used in OrthoDisease is OMIM, which stands for Online Mendelian Inheritance in Man. Anyone can access this database online and search its records at <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM>. This database is especially important for a couple of reasons. The first is that it is a catalogue of all medically interesting human genes. Second, it gives detailed descriptions of each gene,

including its allelic variants, its functions, its inheritance patterns, and so on. Lastly, it lists every major reference that is associated with the particular genes. A diagram of a returned screen in OMIM for the gene responsible for Huntington's disease is found below (figure 4.12). The figure shows only a very small section of the returned screen.



**Figure 4.12.** A screen from OMIM for the gene for Huntington's disease (from Feb. 14<sup>th</sup>, 2006).

**4.10 Extended Cognition in BLAST.** What will now be demonstrated is that uses of BLAST, such as the creation of the databases presented in the previous section, are examples of extended cognition. The homology and orthology databases such as OrthoDisease rely extensively on various computer algorithms. This reliance is necessary, since it would be impossible for any human or group of humans to perform in any timely manner what these computers algorithms perform. The first homology

database by Rubin et. al. (2000), even though simpler than the more reliable orthology database by O'Brien et. al. (2004), is still a sufficient example of the reliance on computer databases and algorithms.

I will now go into the steps that are required for a research program that uses BLAST. The first step is deciding what sequence comparison needs to be made. This decision is made by the scientist, and can involve many different factors, such as which species are to be compared. The scientist may want to compare a newly sequenced molecule to all the species that are found in a database, or to see whether a particular molecule with a known function is found in another species. The search can also involve comparing particular types of molecules, DNA to DNA, DNA to protein, protein to protein, and so on. From this point onwards, I will use the word 'sequence' refer to DNA, RNA or proteins. Lastly, researchers can decide on specifics such as how accurate they wish the search to be, whether they want only results with structural descriptions in the case of proteins, and so on. Thus, the general step performed at this point is:

- 1) Decision on the type of comparison to be made.

Once the decision is made, the database with the required information needs to be found. In many cases, such as those where a researcher is finding homologues to a newly sequenced gene, the researcher would have the sequence information, but this information still needs to be input into a computer in order to be usable by any BLAST algorithm. It is usually the case, however, that all the information a researcher would need would already be in a particular database, such as the GenBank database. The researcher can freely extract that information and then proceed to the next step.

Nonetheless, at this stage, one can generally state that the researcher requires information from particular computer databases:

2) Sequence information from computer databases.

Once the sequences are extracted from the database, then the BLAST algorithm is used by the researchers in order to get the similarity results they need. As shown earlier in the chapter, the algorithm performs numerous steps, which includes finding triplet ‘words’ in the query sequence, finding matches with those triplets, extending the comparison between the two sequences until the ends of the sequences are reached or when the match score goes below a threshold value. This process can be described as generally doing the following:

3) Using the BLAST algorithm.

The sequences that are returned from the BLAST algorithm are all above a particular score. Using that score along with other information about the sequences under comparison, statistical tools are used in order to find the E-values, which is a statistical measure of the likelihood that the returned similarities are possible in a random scenario. So the next step is:

4) Statistical tools used to determine likelihood of sequence similarity.

Lastly, the scientist would analyze the results to determine which conclusions can be drawn, or whether further operations needs to be performed.

5) Analysis of the results.

There is much more information and many more operations that are possible in order to find the particular results that one needs in experiments such as those of Rubin et. al. (2000) and O’Brien et. al. (2004). The general information and operations presented

above, however, are what are found in almost every experiment that uses BLAST. Grouped together, we can make a comparison of the operations performed by computers and those performed by humans, as well as the source of the information used by these operations. This comparison is presented in Table 4.1.

**Table 4.1** Comparison of human and computer information and operations in BLAST.

<b>Computer Databases Used</b>	<b>Programs Used</b>	<b>Human Operations</b>
Sequence information from particular computer databases.	Using the BLAST algorithm.	Decision on the type of comparison to be made.
	Statistical tools used to determine likelihood of sequence similarity.	Analysis of the results

The necessity of computer databases and algorithms in research using BLAST shows that extended cognition is required in order to do this particular kind of research. A human researcher, on his or her own, would not have all the required information and would lack the ability to perform all the operations required.

**4.11 Analogical Reasoning in BLAST.** The purpose of BLAST is not only to find similar sequences, but also to use that similarity to propose further experiments and theories. For example, a researcher does not only use BLAST to find similar sequences, but to determine what type of function a sequence may have based on its similarity to another sequence, or to find model sequences in order to perform various experiments. Rubin et. al.(2000) and O'Brien et. al. (2004) created their databases using BLAST in order for biomedical researchers to find orthologous sequences to use as models for

human sequences, which would, in turn, help in finding treatments for particular ailments. Take, for example, the reaction of a particular drug on a particular sequence in a model species. The analogies that are used in experiments like those are presented in Table 4.2:

**Table 4.2:** Analogical comparisons using BLAST

<b>Sequence from another species</b>	<b>Human Sequence</b>
Sequence <sub>s</sub>	Sequence <sub>h</sub>
Drug <sub>s</sub>	Drug <sub>h</sub>
Drug-result <sub>s</sub>	Drug-result <sub>h</sub>
Interact <sub>s</sub> (sequence <sub>s</sub> , drug <sub>s</sub> )	Interact <sub>h</sub> (sequence <sub>h</sub> , drug <sub>h</sub> )
Cause-result <sub>s</sub> (drug <sub>s</sub> , drug-result <sub>s</sub> )	Cause-result <sub>h</sub> (drug <sub>h</sub> , drug-result <sub>h</sub> )
Cause <sub>s</sub> (interact <sub>s</sub> , cause-result <sub>s</sub> )	Cause <sub>h</sub> (interact <sub>h</sub> , cause-result <sub>h</sub> )

Although ‘BLAST’ is not seen in the table above, it is the tool used to gauge the similarity between the human sequence and the sequence from another species, just as other tools are used to gauge the similarities between the drugs used on both sequences and the result from those drugs. The analogy presented above is almost identical to the ones presented by Shelley (2003), where he argued that animal models were an example of analogical reasoning. The only difference in this case is that table 4.2 uses model sequences instead of model organisms.

Shelley (2003) argued that, although animal models are examples of analogical reasoning, there are always the possibilities of disanalogies. He acknowledges a paper by LaFollette and Shanks (1996) that challenges the use of analogical reasoning with animal models because it is often the case that those models are not scientifically justified. For example, many treatments for particular diseases that work on rats often fail to work on

humans, making them inaccurate in many cases as models of human physiology. Shelley (2003) proposes that the existence of disanalogies arise due to uncovering further information about the species that one is working with rather than any inherent disanalogy in the comparisons being made.

I believe that neither Shelley (2003) nor LaFolette and Shanks (1996) are wrong in their assessments. The physiologies of other species are analogous to that of humans in many cases, and in others they are not. The degree of analogy often depends on specific biomolecular reactions that occur within the various species. For example, the Krebs cycle is required for cellular respiration and is found in all animal species. Any manipulations to that cycle, including manipulations to particular sequences found within that cycle in one species, have the same effect in all species, including humans. However, a drug that manipulates the biomolecular reactions that govern, say, the length of a rat's tail would likely not have the same effect in humans.

The advantage of using BLAST is that it allows biomedical comparisons to be made at the level of biomolecules, since BLAST finds sequences that are similar among various species. If drug  $x$  is hypothesized to have an effect of sequence  $y$  on humans, then it is likely to have the same effect on any sequence that is identical or even similar in another species. Since BLAST gives highly accurate results on the similarity of sequences between species, researchers can be confident that whatever effect a drug has on a sequence from another species will have the same effect on the similar human sequence. This new level of accuracy makes animal models more likely to be analogous, thus more reliable.



This level of accuracy is possible through a combination of scientific methods that I believe is somewhat unique to bioinformatics. I have described how bioinformatics uses extended cognition through computer use and analogical reasoning. However, these methods are also found in many other biological fields, as well as other scientific fields such as physics (see, for example, Humphreys 2004). The combination of these methods is what may be unique to bioinformatics.

The case studies presented above were attempting to find sequences in species that were homologous to those in the human species. As presented above, the purpose of finding these homologies was to make analogical inferences about the human sequences based on findings made on the homologous sequences. Therefore, analogical reasoning is clearly occurring in these cases. Moreover, the method used to find the homologous sequences was extended cognition through computer use. BLAST was the algorithm run on computers to find homologous species, which were used to make analogical inferences. Therefore, there is a new method in bioinformatics where computers are used to make analogical inferences. This new method that combines the two previously used methods is very powerful in generating reliable results. This will be supported further in the 'Epistemic Appraisal' section.

A possible objection to these studies is that, although the sequences under study are similar between the two species, these sequences are not in physiological isolation within the species' bodies. Rats may share sequence  $x$  with humans, for example, but may play a role in different biomolecular mechanisms in both species necessitating further study of the biomolecular mechanisms that occur within the two species. Chapter

6 will present work on finding these mechanisms, and how one can be more confident in the validity of analogical reasoning with animal models.

**4.12 Epistemic Appraisal of BLAST.** The epistemic advantage of using the extended cognition of BLAST research is also apparent when one uses Goldman (1992) and Thagard's (1997) appraisal standards for epistemic practices, which are reliability, power, fecundity, speed, efficiency and explanatory efficacy:

*Reliability:* BLAST should be expected to be an extremely reliable tool due to the precise statistical tools that are employed. A researcher using BLAST is given a measure of statistical warrant with the E-values that are returned with each comparison. Of course, a scientist may require some knowledge about what these statistics mean and their limits, but bioinformatics researchers are always attempting to make these statistical analyses more reliable. What is also interesting about BLAST is that a researcher can manipulate the reliability of the results that are returned from a BLAST search. For example, they can use different scoring matrices, larger 'word sizes', smaller cut-off values, and so on. By making these changes, a researcher would have less results returned, but can be more confident in their accuracy, thus increasing the reliability of the results.

Also, as was seen when comparing the homologous sequence studies by O'Brien (2004) to the animal studies investigated by Shelley (2003), the analogical inferences that are made using BLAST are reliable. This reliability is because there is statistical support for the molecular comparisons being made (the statistics being generated using powerful computational tools) rather than more rudimentary physiological comparisons being made between animals.

*Power:* Due to the fact that BLAST employs algorithms that are designed to compare biological sequences, as well as to produce statistical values that are very precise, BLAST is a powerful tool for generating a large number of scientific results. As was shown in figure 4.9, BLAST gives both similarity scores and E-values in comparing sequences, and presents each pairwise comparison between the input and matched sequences. These results are usually the most reliable for researchers who are interested in finding sequence, structural and functional relationships among sequences. This reliability explains why BLAST is an extremely popular scientific tool, which is used by almost all biologists (Harding, 2005).

*Fecundity:* Because BLAST is readily available over the Internet, many practitioners can generate these large numbers of results. As mentioned already, BLAST is used by most biologists, thus making it likely that BLAST generates truths for many practitioners.

*Speed:* BLAST is a relatively fast algorithm. BLAST results are returned seconds after a researcher inputs a query sequence and defines the parameters of the search. The inputting and definition process is relatively quick as well, due to the fact that researchers can easily extract sequences from databases such as GenBank, and the parameters are easily defined using the drop-down menus available on the BLAST input page (see figure 4.8). A researcher can easily perform many BLAST searches in a day, thus being able to quickly generate a large amount of results.

*Efficiency:* The cost of using BLAST is extremely minimal, and it is probably one of the cheapest procedures available in the scientific world. All a researcher requires in order to perform a BLAST search is a desktop PC or laptop with any kind of Internet

connection. In fact, it would be much more cost-efficient for a researcher to invest in computers that perform bioinformatics tests than in other expensive equipment for wet-lab procedures.

*Explanatory Efficacy:* Because BLAST statistically finds sequences that are similar to one another, it could help to increase the efficacy of explaining the functional relationships between various sequences. For example, a researcher may hypothesize that a particular human gene is implicated in playing a role in a particular disease. BLAST can be used in finding similar genes in other closely related species, and allow researchers to perform experiments relevant to understanding the mechanism associated with that gene. These experiments would thus lend to an understanding of how the human gene functions, and help in explaining the mechanisms involved in the human disease. Databases such as OrthoDisease are designed to find such similar genes, and provide clues for conducting further biological and medical research. Since these databases use extended cognition and analogical reasoning through computer use, the increased explanatory efficacy that is possible with these databases are due to the combination of extended cognition and analogical reasoning through computer use.

**4.13 Summary.** In this chapter, we saw the scientific methods that are present in research that utilizes BLAST. BLAST is a powerful bioinformatics tool that is able to compare biological sequences to entire databases of sequences. BLAST expands on algorithms that perform pairwise sequence alignments by efficiently by repeating these comparisons many times over. Due to the large sizes of the sequence databases, the complexity of the BLAST algorithms, and the statistical analyses that are required on the comparisons,

scientific research that utilizes BLAST can be seen as an example of extended cognition. Also, the inferences using databases that are created using the BLAST algorithm are examples of analogical reasoning, since comparisons are made between sequences whose functions are unknown and sequences that are known. These inferences are examples of a combination of extended cognition through computer use and analogical reasoning, which is a method that may be unique to bioinformatics research. Lastly, the use of BLAST in biological research meets all the standards stated by Goldman and Thagard for the appraisal of an epistemic practice.

## Chapter 5

# Phylogenetic Analyses Case Study

**5.1 Introduction.** Comparing sequence using tools like BLAST has been very useful for many purposes, as was demonstrated in the last chapter. We also saw how BLAST can be extended to compare not only sequences, but also to create entire databases of homologous sequences. One of the most important extensions of sequence aligning algorithms, however, is in the creation of phylogenetic trees. Pairwise alignment algorithms were extended to create algorithms that perform multiple sequence alignments (MSA), and these alignments were created in order to construct phylogenetic trees. Each of these steps will be explained in this chapter.

Phylogenetic trees represent the evolutionary relationships between species, just as family trees represent the relationships between family members. Although phylogenetic trees have been created since at least the time of Carl Linnaeus, bioinformaticists have revolutionized the creation of these trees using genomic and proteomic information instead of the classical methods which relied on physiological, morphological, behavioural and geographical information. The newer tree reconstructions are thus very different from the older ones, and some types of trees, such as those that track human communities, have only become possible with the use of bioinformatics.

This chapter will begin with a short look at the history of the science behind the creation of phylogenetic trees, especially the progress of the science using bioinformatics. Within that history there will be a more detailed description of how phylogenetic trees are

created using bioinformatics tools. Lastly, I will show how these tools utilize the scientific methods of extended cognition through computer use and analogical reasoning, as well as the combination of the two.

**5.2 Classification as Definitions.** The history of phylogeny can be traced to Aristotle's work on definitions and biology. According to Aristotle in his *Topics* (Falcon, 1996), words are defined using a genus and differentium. The genus is the class of objects the word belongs to and the differentium how the word is distinguished from other objects within the class. For example, 'bread' is defined as 'a staple food made from flour or meal mixed with other dry and liquid ingredients, usually combined with a leavening agent, and kneaded, shaped into loaves, and baked'. The genus, in this case, is 'a staple food' and the differentium is what follows in that definition. This definition of 'definition' was extended to Aristotle's classification of biological organisms. In his *History of Animals*, Aristotle created a classification system for organisms that attempted to group animals into broader classes. The most specific instances within a class would be definitions of particular animals. For example, Aristotle defined humans as 'featherless bipeds'. Bipeds, however, belonged to a larger class of animals that were 'live-bearing'. Below is a small portion of Aristotle's classification of animals:

#### I. Blooded Animals

##### A. Live bearing animals

###### 1. Homo Sapiens

###### 2. Other mammals without a distinction for primates

##### B. Egg-laying animals

1. Birds

2. Fish

II. Non-Blooded Animals

A. Shell skinned sea animals: Testacea

B. Soft shelled sea animals: Crustacea

C. Non-shelled soft skinned sea animals: Cephalopods

D. Insects

E. Bees

III. Dualizers (animals that share properties of more than one group)

A. Whales, seals and porpoises—they give live birth yet they live in the sea

B. Bats—they have four appendages yet they fly

C. Sponges—they act like both plants and animals

(From Aristotle's *The History of Animals*, translated by Balme, 1991)

Biological classifications, since Aristotle's work, continued on a trend of classification based on visually salient features, which included whether animals were found in water or were domesticated. However, due to the vague nature of distinguishing between animals and delineating the classes, these classifications were usually somewhat arbitrary. For example, some defined fish as animals that lived in water, and as such hippopotami were classified as fish.

The next major advancement in biological classifications came with Carl Linnaeus in the 18<sup>th</sup> century. According to his *Systema Naturae*, his classification of organisms consisted of 5 levels: species, genera, orders, classes and kingdoms. His

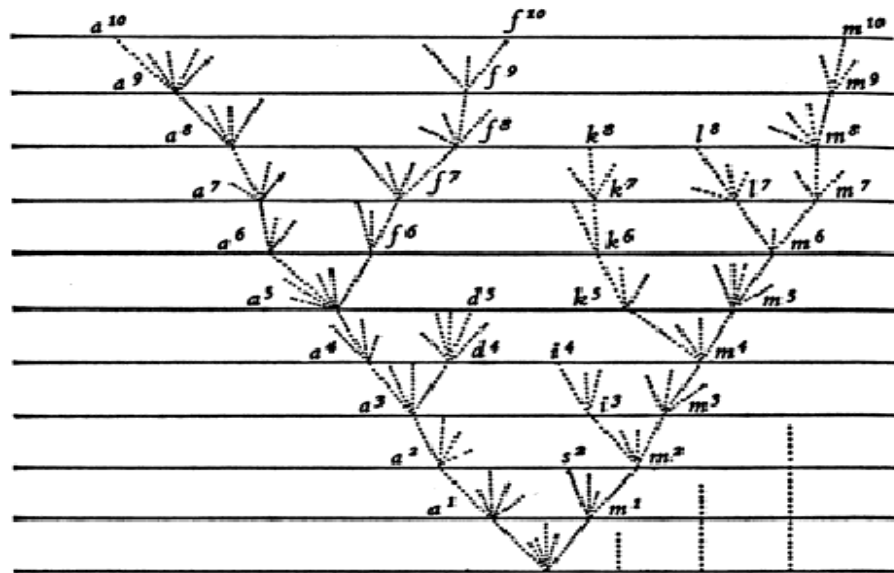


classification system proved to be much more precise than Aristotle's since he was the first to distinguish species based on their sexual behaviour and characteristics (Ridley, 1996).

**5.3 Darwin's Theory of Evolution by Natural Selection.** The next step in making species classifications more precise came with Darwin's work on evolution by natural selection. This work was important for classification since it gave an underlying theory of why organisms shared similarities to each other in varying degrees. That theory is that all species evolved from common ancestors. The degree of relationship between species depends upon when those species diverged from a common ancestor. This degree of divergence is an important consideration when classifying species.

Since Darwin's addition of evolution to biological classification, *phylogenetic trees* have been used in order to represent the evolutionary relationships between species. Figure 5.1 is one of the first of such trees, which was drawn by Darwin in his *Origin of Species* (1859). The root of the tree represents the first species. The branching out of the tree represents the evolution of new species from the original species. The growth of the tree represents time, with the root being the oldest species and the top of the tree representing extant species. The branches that terminate before reaching the top of the tree represent species that have gone extinct. In this figure, there are two main branches that represent two separate *kingdoms*. Darwin only distinguished two kingdoms in his day: plants and animals. The branch on the left splits into two more major branches, which would represent two separate *classes*. If these latter branches are within the plant kingdom, then it may represent a division between trees and non-woody plants.

With Darwin's work, the classification of species was no longer as arbitrary as in earlier biological work. Biologists no longer classified species based on the greatest number of similarities, but classified according to similarities that one assumed were more or less affected by the forces of evolution. For example, modern biologists believe that evolving from a quadruped to a biped does not require many evolutionary changes, as all that is required are some changes in bone-structure and neurological modifications for balance. Therefore, species that are relatively similar except for the way they move on land are likely to be closely related. On the other hand, the outer covering of an animal (scales, feathers or furs) require many evolutionary changes in order for one to evolve into another. Therefore, any species that are relatively similar except for their outer covering are still likely to be distantly related.



**Figure 5.1** Darwin's tree of life from his *Origin of Species* (1859, image is in the public domain).

Thus Darwin's theory of evolution by natural selection gave biologists a better basis for how to classify organisms. Despite this basis, however, many questions still remain. One major problem is in determining the actual rates that particular traits evolve. This information is important in the creation of trees since it gives a better representation of where the nodes on the trees should be.

Recalling the examples presented above, we assumed that bipedalism was easier to evolve for a quadruped than it was for a species to evolve fur when it was feathered. This 'relative comparison' assumption is generally easy to accept based on basic knowledge of animal physiology. However, another problem arises when, for many traits, it seems extremely difficult to assess their rate of evolution. For example, it is still relatively difficult to determine the rate of the evolution of certain morphological traits, say from legs in a reptile to wings on a bird. Another example of this difficulty is in figuring out the ease by which particular behavioural traits arise. A specific example is our general lack of knowledge of the evolution of cognitive abilities, where we still are not sure how much 'evolution' would be required for, say, a monogamous breeding behaviour to evolve from one that is polygamous. The ease of evolution of cognitive abilities is a debated topic in philosophy of mind and cognitive science (see, for example, Quartz & Sejnowski, 1997 and Pinker, 2003)

Another problem that phylogeneticists encounter is in determining the exact time-span needed for particular evolutionary events to occur. Looking only at extant species does not adequately help with this, so biologists have turned instead to the fossil record. The difference in age between the fossils of different species gives a very rough estimate of the time needed for particular traits to evolve. However, the fossil evidence will

always be incomplete, and is only available for certain species, specifically those that either have shells or an internal skeleton.

**5.4 Molecular Turn.** From the development of heredity studies starting with Mendel in 1866, and also from work on molecular biology starting with Pauling (1939), new and more precise methods arose for creating phylogenetic trees. These molecular investigations allowed researchers to compare the molecular sequences among species, and be able to measure differences between species much more precisely.

The first researcher to make such a comparison was Frederick Sanger, who sequenced the insulin protein in 1953 (Alberts, 2002). Sanger sequenced this protein from 5 species: cow, sheep pig, horse and whale. The insulin protein has three sections: A, B and C chains. When Sanger and others compared the A and B chains of the 5 species, they found that the molecular structure was generally highly conserved. The only significant differences were found in the A chain's disulphide 'loop' and in the C chain. It was later discovered that the C chain could undergo many changes without significantly changing the function of the protein. Due to the unhindered evolution of the C chain, researchers could give a more precise estimate on the evolutionary relatedness among species. For example, if there is a 50% similarity between, say, the cow and sheep C chain whereas there is a 75% similarity between the cow and horse C chain, then it is more likely, all else being equal, that the horse and cow are more closely related than the sheep and cow.

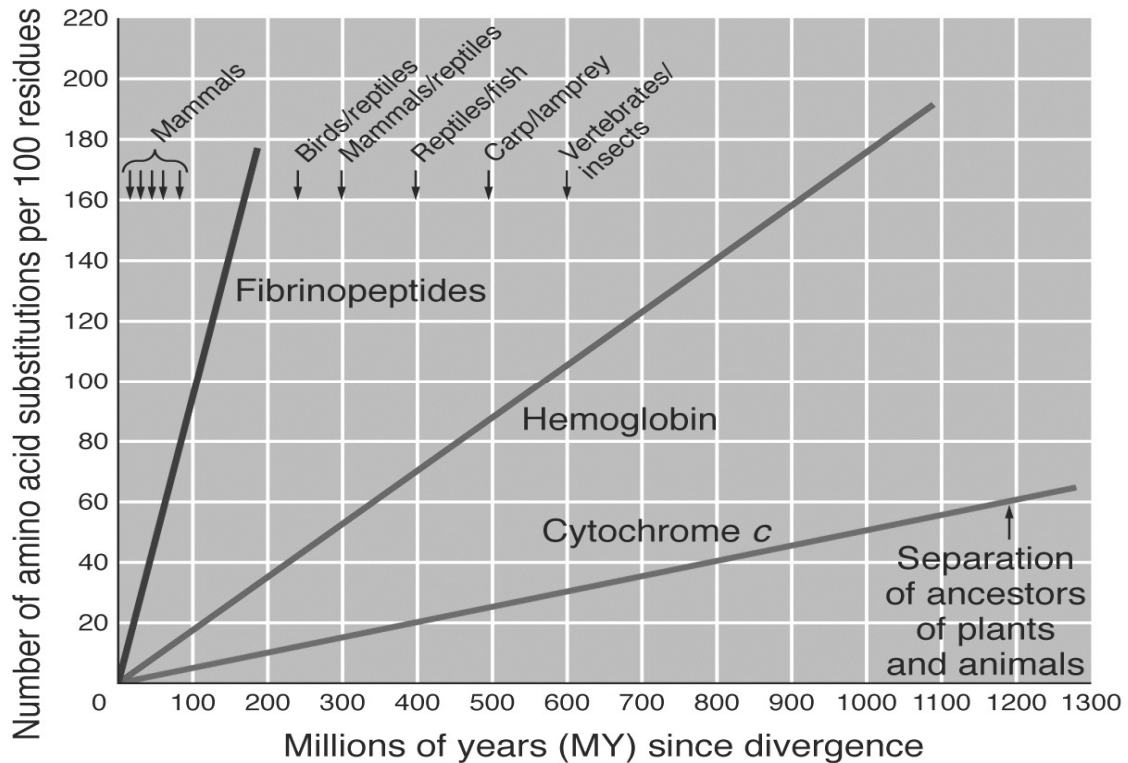
Another interesting phenomenon that was discovered with the analysis of biological molecules is the seemingly non-linear relationship between molecular changes

and phenotypic changes. For example, changes in the disulfide loop in chain A does not result in any phenotypic changes. However, Perutz and Kendrew (1965) found that changes in only a couple of amino acids in hemoglobin resulted in very significant phenotypic changes. Due to this discrepancy between genotypic and phenotypic changes, some researchers hypothesized that there may exist some kind of 'molecular clock', and that these clocks were different for each gene, depending upon its phenotypic stringency. These clocks can contribute to creating relatively accurate estimates of the times of divergence between species, since the number of genotypic changes in particular genes would correspond to times of divergence between species.

Richard Dickerson (1971) found evidence for such a clock. He used three proteins, cytochrome *c*, hemoglobin, and fibrinopeptides, for which there was abundant sequence information for a number of species. He plotted the number of changes in the peptides versus the previously hypothesized (using paleontological data) number of years that the species had diverged. Dickerson found that all the plots lay on a straight line, which supported the molecular clock hypothesis. What was also interesting was that each of the protein families had different rates of change. Further analysis into the functions of these proteins showed that the rate of change may be dependant upon the functional constraints on each protein as dictated by natural selection. In this case, fibrinopeptides are not as strictly constrained as cytochrome *c*, allowing the former to evolve much more quickly. The results are shown in figure 5.2.

Further studies have provided some exceptions to the molecular clock hypothesis, but it has since been found to generally hold (Pevsner, 2003). What this relative

constancy means is that researchers not only have a better method for measuring differences between species, but also have a better measure for *when* species diverge.



**Figure 5.2** Plot of evolving peptides supporting the ‘molecular clock’ hypothesis. (from Griffiths et. al., 2002)

**5.5 Import from Sequence Comparison Algorithms.** The methods and discoveries described above were the developments in phylogenetics that preceded bioinformatics. The first bioinformatics tools that could be useful for phylogenetic studies were, of course, the same tools that were useful in the BLAST case, namely, pairwise comparison algorithms. Since these algorithms are used to compare sequences, and phylogenetic studies now involve the comparison of molecular sequences, phylogenetic studies will

certainly benefit from these pairwise comparison algorithms. The difference between phylogenetic studies and other pairwise comparisons, however, is that phylogenetic studies do not only compare one sequence to another but may involve the comparison of many sequences at once. What is also required, in this case, is the creation of new tools that can create the actual trees based on the output of these comparisons. The ‘distance-method’ of tree creation, which I will describe later in the chapter, does not require this comparison.

Dayhoff et al. (1978) were the first to provide new tools for phylogenetic studies. As presented in the last chapter, Dayhoff created the PAM (Point Accepted Mutation) matrix that gave scores for changes in particular amino acids. This was important in order to provide an accurate measurement of how similar proteins are. For example, threonine changing to valine has a score of  $-6$  whereas threonine changing to tryptophan has a score of  $-19$ . This means that sequences that differ in the former case are not as different as those that differ in the latter case.

Another important tool is the Needleman-Wunsch pairwise comparison algorithm (1972). This tool was described in the last chapter as an efficient means for comparing two sequences. Instead of simply checking each unit with every other unit, this algorithm gives scores to particular matchings, thus giving an overall similarity score between matched sequences. Another benefit of this algorithm was its ability to take ‘gaps’ into account, which is an important issue when dealing with biological sequences.

If a researcher wished to create a phylogenetic tree, they could simply use this latter algorithm by comparing the scores of each comparison. However, they would be presented with a major problem, which is that of not truly finding an evolutionary pattern

when comparing the similarities and differences. Consider the following example, where sequences from the following species are known: human, chimpanzee, cow, chicken and dog. One can compare each of these sequences individually to discover their relative similarities. Let us use the beta-globin sequence from each species. Below is the FASTA format for the beta globin protein from each species.

```
Human: >gi|4504349|ref|NP_000509.1| beta globin [Homo sapiens]
MVHLTPEEKSAVTALWGKVNVDDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMG
NPKVKAHGKKVLGAFSDGLAHLNLDKGTFAATLSELHCDKLHVDPENFRLLGNVLCVLA
AHHFGKEFTPPVQAAAYQKVVAGVANALAHKYH
```

```
Chimpanzee: >gi|38227|emb|CAA26204.1| beta-globin [Pan troglodytes]
MVHLTPEEKSAVTALWGKVNVDDEVGGEALGRLLVSRLLVVYPWTQRFFESFGDLSTPDA
VMGNPKVKAHGKKVLGAFSDGLAHLNLDKGTFAATLSELHCDKLHVDPENFRLLGNVLCV
LAHHFGK
```

```
Cow: >gi|395|emb|CAA25111.1| beta-globin [Bos taurus]
MLTAEKAAVTAFAWGKVKVDEVGGEALGRLLVVYPWTQRFFESFGDLSTADAVMNNP
KVKAHGKKVLDSEFSNGMKHLDDLKGTFAALSELHCDKLHVDPENFKLLGNLVVLA
RNFSGKEFTPVLAQDFQKVVAGVANALAHRYH
```

```
Chicken: >gi|408500|gb|AAD03346.1| beta-H globin [Gallus gallus]
MVHWTAEKQLITGLWGKVNVAECGAEALARLLIVYPWTQRFFASFGNLSSATAIIGNP
MVRAGKVKVLDSEFSGEAVKNLDNIKKSFAQLSKLHCDKLHVDPENFRLLGDILIIVLASHF
SKDFTPASQAAWQKMVRVVAHALAHEYH
```

```
Dog: >gi|57113367|ref|XP_537902.1| PREDICTED: similar to beta globin [Canis
familiaris]
MVHLTAEKSLVSGLWGKVNVDDEVGGEALGRLLIVYPWTQRFFDSFGDLSTPDAVMSN
AKVKAHGKKVLDSEFSGLKNDNLKGTFAKLSELHCDKLHVDPENFKLLGNVLCVLA
HHFGKEFTPPVQAAAYQKVVAGVANALAHKYH
```

The next step would be to compare each of these sequences with each other. This comparison would generate a similarity score between each species. Based on that score, a researcher can get an estimate about the evolutionary relationships between these species. The following table (Table 5.1) contains the similarity scores and E-values for each of the matched sequences using the Smith-Waterman algorithm. This simple tool is



available for public use as a web application on the NCBI site at the following web page: [www.ncbi.nlm.nih.gov/blast/bl2seq/wblast2.cgi](http://www.ncbi.nlm.nih.gov/blast/bl2seq/wblast2.cgi). The reader must keep in mind, however, that this stage is not what is performed in current research where phylogenetic trees are created using bioinformatics tools, since there are some necessary steps missing, and what is actually used in bioinformatics research is a completely different algorithm. This faulty method is presented here in order to help the reader better follow the upcoming sections.

**Table 5.1.** Scores and E-values of beta-globin sequences from various species.

<b>Comparison</b>	<b>Score</b>	<b>E-value</b>
Human:Chimpanzee	246	2e-64
Human:Cow	252	4e-66
Human:Chicken	212	3e-54
Human:Dog	275	43-73
Chimpanzee:Cow	203	1e-51
Chimpanzee:Chicken	173	2e-42
Chimpanzee:Dog	220	1e-56
Cow:Chicken	199	2e-50
Cow:Dog	247	1e-64
Chicken:Dog	301	5e-81

Once the comparisons are made, the similarity scores are weighed against each other in order to determine the degree of relationship between the sequences. Constructing a tree at this point becomes a problem, however, since it is not altogether clear where to begin. One can start at the sequences that had the highest score, and make them the closest relatives on the tree. In this case, the highest score is between the dog and the chicken (score = 301, E-value = 5e-81). This result is surprising, since we would expect the dog sequence to be more closely related to the other mammals. Alternatively, one can start with the sequences with the lowest matching score and begin the major branching from

there. In this case, the lowest score was between the chicken and the chimpanzee (score = 173, 2e-42). In this case, the latter strategy seems the most accurate, but we will suspend judgment on either strategy for now, since these two strategies will be discussed after we look at multiple sequence alignments, which is the major tool used by bioinformaticists in constructing phylogenetic trees.

**5.6 Multiple Sequence Alignments.** Chapter 4 described how the Smith-Waterman algorithm was extended into the BLAST tool. This powerful tool was able to perform pairwise comparisons at an even faster rate than previous algorithms, which allowed for sequences to be run against massive databases in relatively short times. Feng and Doolittle (1987), however, used the algorithm for a different purpose. They extended it to be able to perform what they termed “multiple sequence alignments”, or MSA for short. MSAs are able to compare multiple sequences at once, rather than running multiple pairwise alignments. In comparing multiple sequences with one another, one can certainly perform individual comparisons and compile the similarity scores afterwards to determine the relationship between all the sequences, as attempted in the beta-globin example above. However, the individual comparisons do not take into account the ways that these sequences are all similar or different. For example, consider the following four fictional sequences:

1: ABCDEFG

2: ABCDEGH

3: ABCDEHI

4: AJCDEFG

In this comparison, the first three sequences have a difference of two residues at the end of the sequence: FG, GH or HI. The rest of the sequence, however, is identical. Sequence 4, on the other hand, has only one residue difference with sequence 1: it has a J instead of a B. As a result, it seems likely that sequences 1 and 4 are more closely related than they are to the other two sequences. However, due to the fact that residues A through E are conserved in sequences 1 to 3, then it is likely that that sequence of residues has an important function, and is generally evolutionarily conserved. Due to this possible importance, it becomes more likely that sequence 1, 2 and 3 are more closely related than any of them are to sequence 4. The MSA algorithm of many phylogenetic tree-creation applications is able to find these highly conserved areas, which are called *domains*, and take them into account when determining the similarities among a number of sequences.




Specific programs are used to create multiple sequence alignments. The most popular program is ClustalW (Thompson et. al., 1994), which can be accessed online at <http://www.ebi.ac.uk/clustalw/>. Sequences are inputted in their FASTA formats, and after specifying a few other parameters, such as gap penalties and word sizes, a researcher can obtain a multiple sequence alignment which includes the following: the alignment, the regions which are most likely to be conserved, the similarity scores between the sequences, as well as a phylogenetic tree generated from the MSA. I will not go into how the tree is generated until in the next section. Figure 5.4 shows the input screen of ClustalW and figure 5.5 is the output screen using the beta-globin sequences from our original example, however the tree generated from this output is not displayed in this case, since the figure would take up more than a page.

**EMBL-EBI**  
European Bioinformatics Institute

EBI Home About EBI Groups Services **Toolbox** Databases Downloads Submissions  
SEQUENCE ANALYSIS

**ClustalW Submission Form**

Clustal W is a general purpose multiple sequence alignment program for DNA or proteins. It produces biologically meaningful multiple sequence alignments of divergent sequences. It calculates the best match for the selected sequences, and lines them up so that the identities, similarities and differences can be seen. Evolutionary relationships can be seen via viewing Cladograms or Phylograms. [New users, please read the FAQ.](#)

>> [Download Software](#)   

YOUR EMAIL	ALIGNMENT TITLE	RESULTS	ALIGNMENT	CPU MODE
<input type="text"/>	Sequence	interactive ▾	full ▾	single ▾
KTUP (WORD SIZE)	WINDOW LENGTH	SCORE TYPE	TOPDIAG	PAIRGAP
def ▾	def ▾	percent ▾	def ▾	def ▾
MATRIX	GAP OPEN	END GAPS	GAP EXTENSION	GAP DISTANCES
def ▾	def ▾	def ▾	def ▾	def ▾

OUTPUT		PHYLOGENETIC TREE		
OUTPUT FORMAT	OUTPUT ORDER	TREE TYPE	CORRECT DIST.	IGNORE GAPS
aln w/numbers ▾	aligned ▾	none ▾	off ▾	off ▾

Enter or Paste a set of Sequences in any supported format: [Help](#)

Upload a file:

**Figure 5.3** ClustalW input screen from <http://www.ebi.ac.uk/clustalw/>, from March 19<sup>th</sup>, 2006. A researcher can set the parameters using the drop-down menus and input the FASTA formatted sequences in the input window.



Returning to our original example of comparing globin sequences from various species, figure 5.5 has the new alignment scores using the MSA. Table 5.2 compares these scores with the ones found using simple pairwise alignments.

**Table 5.2** Comparison of MSA scores with pairwise alignment scores

<b>Comparison</b>	<b>Pairwise Score</b>	<b>MSA Score (% ID)</b>
Human:Chimpanzee	246	96
Human:Cow	252	84
Human:Chicken	212	67
Human:Dog	275	89
Chimpanzee:Cow	203	80
Chimpanzee:Chicken	173	65
Chimpanzee:Dog	220	85
Cow:Chicken	199	64
Cow:Dog	247	82
Chicken:Dog	301	69

The pairwise alignments generated some results that were somewhat unexpected. The highest score was between the dog and chicken sequences, and this is strange given that we assume that dogs are more closely related to other mammals than to birds (maybe domesticated animals have their own evolutionary branch after all). The MSA alignments, on the other hand, appear more accurate. The highest score is between the human and the chimpanzee, species that are also known, based on current research, to be the closest related species in this list. The lowest score is between the cow and chicken, but the human and chimpanzee alignments with the chicken sequence are almost as low.

Looking at the MSA that was generated (lower half of figure 5.4), we see that these new scores were due to the algorithm's ability to find the conserved domains. The regions with domains are the ones with the '\*' at the bottom of the MSA, indicating that

the residues in these positions are identical in all the sequences. The positions with the double-dots at the bottom mean that the residues are somewhat similar, whereas those with the single dots are only slightly similar. Any deviations from those domains would create a large penalty in the similarity score, and would thus mean that the species from which those sequences originated would be more distantly related to the other species in the comparison.

**5.7 Tree-Building.** As mentioned earlier, there are two main methods for building phylogenetic trees once similarity scores have been generated (Hall, 2001): distance-based methods, and character-based methods. Both use information from the multiple sequence alignments in different ways.

Distance-based methods calculate the degrees of divergence. A simple measure that calculates the degree of divergence is called the Hamming distance (Hamming, 1950):

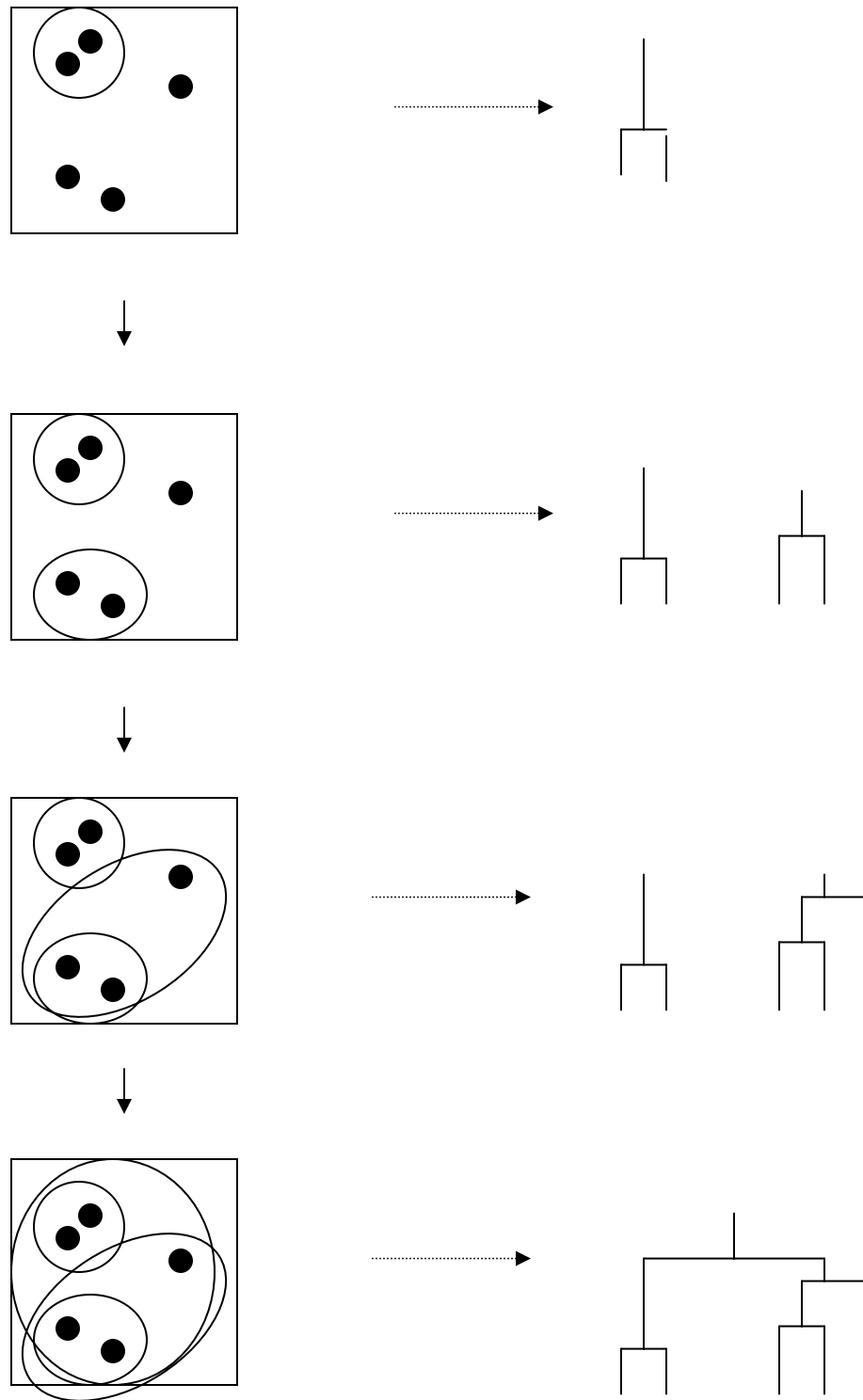
$$D = n/N \times 100,$$

where  $D$  is the degree of divergence,  $N$  is the number of residues in the sequence and  $n$  is the number of sites at which there are differences (Pevsner, 2003, p.378). The sequences that are most alike, meaning those with the lowest Hamming distances, are grouped together first. After these first two sequences are grouped, new similarity scores are calculated between the created group and the remaining sequences. This is done until all the sequences are grouped together. These calculations give indications on the branch lengths, which are supposed to represent the evolutionary diversions between the species. The steps for creating a tree using one type of algorithm grouped that is classified into

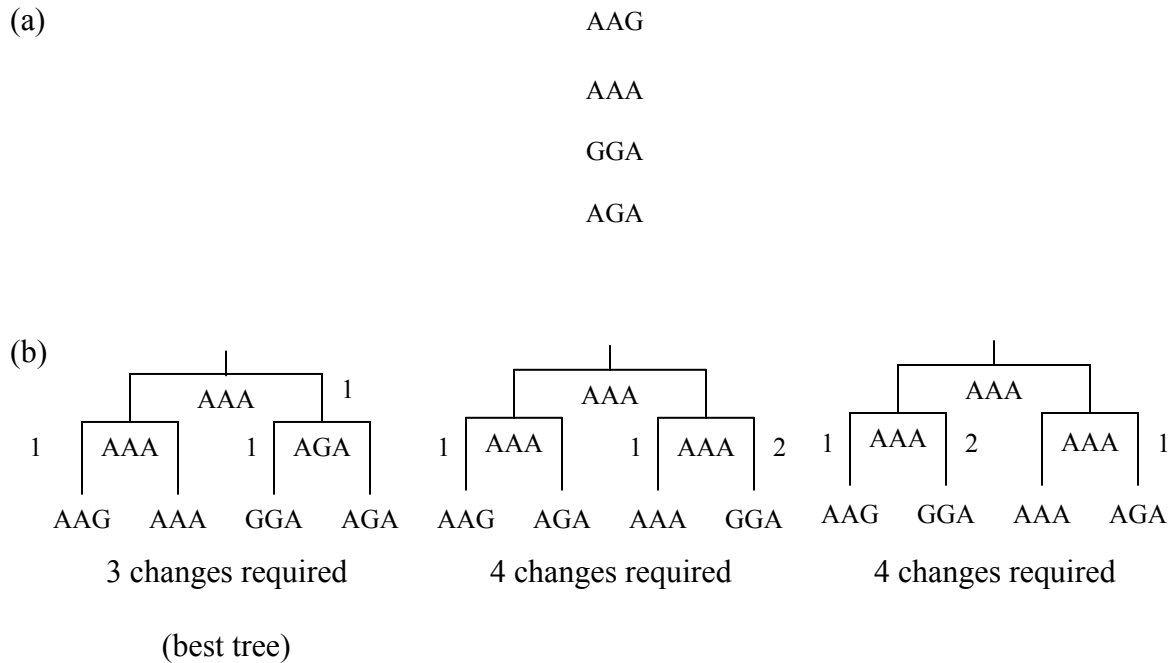
this method are represented in figure 5.6. This algorithm is called the unweighted pair group method with arithmetic mean, or UPGMA for short (Sokal & Michener, 1958). The dots in the boxes on the left represent the sequences and the distances between the dots represent the degree of divergence. The trees on the right represent the tree building process. Each of these dots is put into a group depending on its distance from others, until all the dots are part of at least one group. The final tree in the diagram below is the final phylogenetic tree created after using this distance-based method.

Character-based methods, on the other hand, analyze alignments in order to find the most parsimonious tree, which is the one that would involve the least number of amino acid changes for each branch. Figure 5.7 shows a number of sequences and the possible trees that can be created using those sequences (in the example, 3 trees need to be built, but only 2 are shown). In this case, the first tree is selected since it requires the least number of amino acid changes. This is an example of the Maximum Parsimony, or MP algorithm, which is classified as a character-based method (Eck & Dayhoff, 1966, Fitch, 1977).



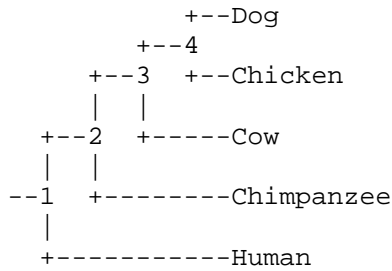


**Figure 5.5** Representation of Distance-based method, using the UPGMA algorithm, of tree building (from Pevsner, 2003, p. 382).



**Figure 5.6** Principle of character-based tree building. The tree with the least number of amino acid changes (the one on the left) is selected as the best tree (from Pevsner, 2003, p.385).

Any scientist, through the use of various websites, can input sequences from any number of species and run any type of tree-creating algorithm they wish. Popular examples of web sites that offer programs that anyone can download are PAUP ([paup.csit.fsu.edu](http://paup.csit.fsu.edu)) and Mr. Bayes ([mrbayes.csit.fsu.edu](http://mrbayes.csit.fsu.edu)), and WebPhylip allows anyone to run these different algorithms while online ([biocore.unl.edu/WEBPHYLLIP](http://biocore.unl.edu/WEBPHYLLIP)). Returning to the example above (table 5.2), below (figure 5.8) is the constructed tree using the beta-globin alignment, where the MSA was done using the ClustalW website, and an MP tree was constructed using the WebPhylip website.

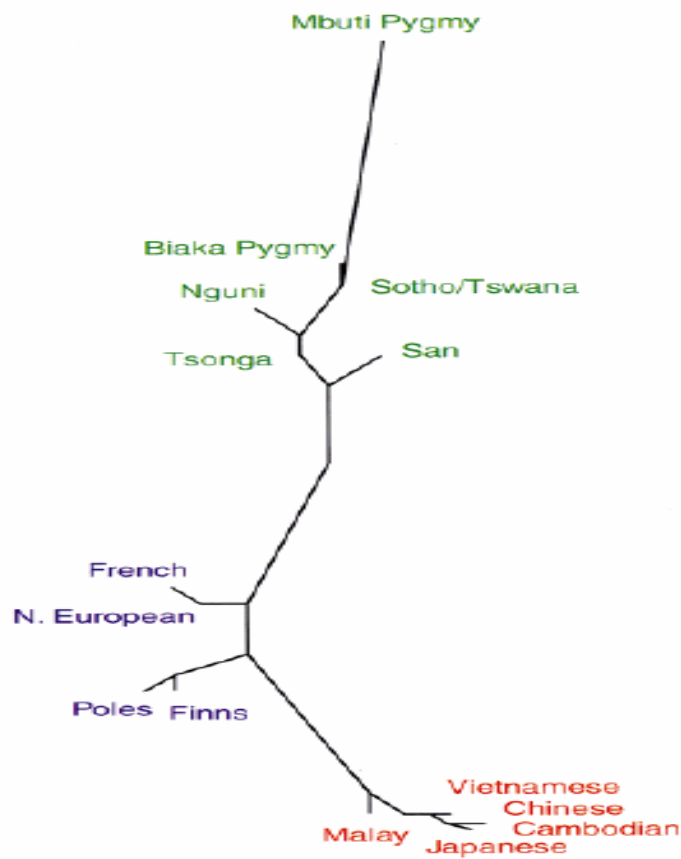


**Figure 5.7** Phylogenetic tree created from a MSA of beta-globins from various species, using a maximum-parsimony algorithm to create the tree.

**5.8 Phylogenetic Tree Case: Human Evolution.** One of the most interesting questions for evolutionary researchers concerns the evolution of the human species. Some questions raised are: what is the age of the *Homo sapiens* species, how did it originate, and what were its subsequent genetic changes. Until this century, most scientific analyses of human evolution were accomplished using human fossils. This type of study is still quite popular, although it is slowly being replaced by genetic research. The first researchers to use genetic markers in determining human evolution were Rebecca Cann et. al. (1987). Since information on the human genome was still very limited during the time of their research, Cann et. al. used mitochondrial DNA (DNA found in human mitochondria). This DNA is inherited maternally, and evolves several times faster than regular, cellular DNA, and is a smaller genome (16kB versus 3GB). Due to this difference, mitochondrial DNA has some advantages over cellular DNA, since recombination of cellular DNA makes it more difficult to trace genetic history, and the higher rates of mutation in mitochondrial DNA allow for one to investigate shorter periods of evolutionary history.

Cann et. al., however, did not use any alignment software or tree-building algorithms, thus their use of bioinformatics tools was somewhat minimal. Subsequent studies have used more modern bioinformatics tools. Two theories on the origin of modern humans have been proposed since the use of genetic data became common. The more popular “Out-of-Africa” theory (Nei 1995, Stoneking 1994, Bowcock et. al. 1994) posits that modern humans originated in Africa and replaced older hominid species across the globe. This origin would explain the great genetic diversity that is observed among African populations compared to the lesser diversity among non-African populations. The less popular “Multi-regional” theory holds that modern humans evolved independently from various groups across the globe, and that the genetic homogeneity that exists outside the African populations is due to natural selection and genetic exchange among those populations. Although humans are spread across the globe and look very different, we, including African populations, are actually relatively genetically homogeneous. Chimpanzees have more genetic variation amongst themselves than do humans (Crouau-Roy et. al., 1996). This evidence could only be found by investigating the genetic data.

Multiple-sequence alignments of different DNA markers as well as tree constructions using character-based methods have generated trees such as the one in figure 5.9 (from Jorde et. al., 1998). This figure shows the relative diversity of African populations (the populations found at the top of the tree) compared to other human populations. The tree in figure 5.9 is unrooted, which means that there is no conjecture about the origin of all these populations, although most researchers agree that the human species originated in Africa.



**Figure 5.8** Unrooted phylogenetic tree of human genetic relations (from Jorde et. al, 1998)

**5.9 Extended Cognition in Phylogenetic Tree Creation.** Using the same strategy from the BLAST case study, I will show how the creation of phylogenetic trees using bioinformatics tools is a case of extended cognition. As explained in Chapter 2, extended cognition in bioinformatics includes representations in computer databases and representations developed using computer algorithms. Similar to BLAST, phylogenetic tree creation requires a researcher to gather information on the sequences as well as to use the algorithms necessary to construct such trees. Specific to this case are the

particular multiple sequence alignments that are performed as well as the creation of the trees themselves. The multiple sequence alignment algorithms involve many computational steps, from determining the probability that particular residues are found in particular positions, to determining the probabilities that particular sequences arise, to the creation of similarity scores. The tree creation, which can be done using either a distance-based or character-based algorithm, involves using those similarity scores in order to find the most closely related species as well as determining the length of each branch, which represents the genetic similarities between the species.

The steps for creating phylogenetic trees will now be investigated. The first step in creating a phylogenetic tree is deciding which species to compare. Depending on the study being performed, the species to be compared can hypothetically be closely related or distantly related. One example of the former case includes the comparison of human populations. An example of the latter case may be the comparison of bird species with reptile species, in trying to determine when those two lineages may have diverged. The general step made at this point is:

- 1) Decision on species that are to be part of phylogenetic tree.

The researcher also needs to make other decisions based on previous theories and results. One of those decisions is what genes from the chosen species to use. In the example above, we chose the beta-globin gene. Another set of decisions is what algorithms to use. These decisions are often based on what mathematical assumptions the researcher decides are acceptable.

- 2) Decision of what genes and algorithms to use.

Once this decision is made, the genetic or protein sequences of the species need to be compiled either through sequencing experiments performed by the researchers, or extracted from particular databases. This latter option is similar to the one presented in the BLAST chapter, where one can find the sequence data by visiting the NCBI web site and requesting particular sequences, such as, human beta-globin. So the third step is:

- 3) Sequence information from particular computer databases.

Once all the sequence data is compiled, according to the method I presented above, the next step is to perform a multiple sequence alignment on these sequences. One can use the ClustalW algorithm, which is publicly available at <http://www.ebi.ac.uk/clustalw/>, and input the FASTA format of the sequences in the input window shown in figure 5.4. These programs return the alignment of these multiple sequences, as well as distance scores, which are measures of how closely related each sequence is to the others. Therefore, step four is:

- 4) Use of multiple-sequence alignment algorithms.

After the distance scores are calculated using the multiple-sequence alignment, the phylogenetic trees can be generated using distance-based or character based methods. There are web sites that perform this step separately, as well as offering many different types of tree-building software. For example, the T-Coffee web site runs the ClustalW analysis (Thompson et. al., 2004), and these results can be inputted into a tree-building program such as MrBayes (Huelsenbeck & Ronquist, 2001). As presented in figure 5.8, the ClustalW web site performs the multiple sequence alignment and creates the trees as well. Therefore, step five is:

- 5) Use of tree-construction algorithms

Lastly, as in the BLAST case, the scientist would analyze the results to determine which conclusions can be drawn, or whether further operations needs to be performed.

6) Analysis of the results

This is a brief overview of the steps that are likely taken by a researcher in order to produce a phylogenetic tree using bioinformatics tools. Generally, as in the case of BLAST, the main role for the scientist simply involves choosing the type of comparison to be made, possibly the type of algorithm to use in order to generate the tree, and an analysis of the results. Otherwise, all other steps require the use of a computer database or algorithm, many of which are publicly available on the Internet. Table 5.3 shows a comparison between the human and computer information and operations.

**Table 5.3** Comparison of human and computer information and operations in the creation of phylogenetic trees.

<b>Computer Databases Used</b>	<b>Programs Used</b>	<b>Human Operations</b>
Sequence information from particular computer databases	Use of multiple-sequence alignment algorithms	Decision on species that are to be part of phylogenetic tree
	Use of tree-construction algorithms	Decisions on what genes and algorithms to use Analysis of the results

The comparison above shows that using bioinformatics for creating phylogenetic trees is certainly a process that involves extended cognition. The human researcher is involved in making the initial decisions and final analyses, but all information and actual sequence alignment and tree creation are performed entirely using computer algorithms. What is even more interesting, as was also shown in my example using beta-globin, is



that the creation of these trees has become a relatively quick and easy process, much easier than the process used by earlier researchers. Although there are many mathematical and evolutionary assumptions made when using these algorithms, it does seem that researchers are finding the use of these algorithms more trustworthy than using previous methods due to the ability to accurately compare molecular sequences.

**5.10 Analogical Reasoning in Phylogenetic Tree Creation.** The investigation into analogical reasoning used in the creation of phylogenetic trees is similar to how it is used in BLAST in that it is found in only some cases. Although sequences are compared to one another in the creation of a phylogenetic tree, this creation is not necessarily an example of analogical reasoning. Consider the example that I used above with beta-globin. Sequence comparisons are required in order to create the phylogenetic tree, but there are no ‘known’ sequences in this case, which are required in order to have analogical comparisons. Since all the sequences and phylogenetic relationships are all relatively unknown in most examples of phylogenetic tree creation, then we cannot claim that these represent true cases of analogical reasoning.

However, there is at least one case of the creation of phylogenetic trees that does utilize analogical reasoning. This is the case where a tree has already been created, and a new species is compared to the members on that existing tree to determine which species it is most closely related to. This is a very common type of research study, and there are large projects that have extensive phylogenetic trees and are in the process of adding more species to that tree. One such project is the “Tree of Life Web Project”, which is found at <http://tolweb.org/tree/> (figures 5.10, 5.11). The main page of the site states:

The Tree of Life Web Project (ToL) is a collaborative effort of biologists from around the world. On more than 4000 World Wide Web pages, the project provides information about the diversity of organisms on Earth, their evolutionary history (phylogeny), and characteristics.

The site contains a comprehensive phylogenetic tree, and allows researchers to add new species to the tree or to revise the phylogenetic mappings. The simple additions to this tree would involve analogical reasoning since the new addition is being compared to the species that are already found on the tree. Table 5.4 presents the analogical reasoning behind a particular addition that is made to a phylogenetic tree using beta-globin sequences from both the unknown and known species.

**TREE OF LIFE web project**

Explore the Tree of Life

**Browse the Site**

- [Root of the Tree](#)
- [Popular Pages](#)
- [Sample Pages](#)
- [Recent Additions](#)
- [Random Page](#)
- [Treehouses](#)
- [Biographies](#)
- [Picture Sampler](#)

[Search](#)  
[advanced](#)

---

**News**

Each revision of a Tree of Life branch or leaf page is now archived...

[read more](#)

**Learn about ...**

**Antennariidae**  
(frogfishes)

[image info](#)

"Frogfishes of the family Antennariidae are typically small, globose anglerfishes..."

[read more](#)

[previously featured pages](#)

---

The Tree of Life Web Project (ToL) is a collaborative effort of [biologists from around the world](#). On more than 4000 World Wide Web pages, the project provides information about the diversity of organisms on Earth, their evolutionary history ([phylogeny](#)), and characteristics.

Each page contains information about a particular group of organisms (e.g., [echinoderms](#), [tyrannosaurs](#), [phlox flowers](#), [cephalopods](#), [club fungi](#), or the [salamanderfish of Western Australia](#)). ToL pages are linked one to another hierarchically, in the form of the evolutionary tree of life. Starting with the root of all Life on Earth and moving out along diverging branches to individual species, the [structure of the ToL project](#) thus illustrates the genetic connections between all living things.

[read more about the Tree of Life Web Project...](#)

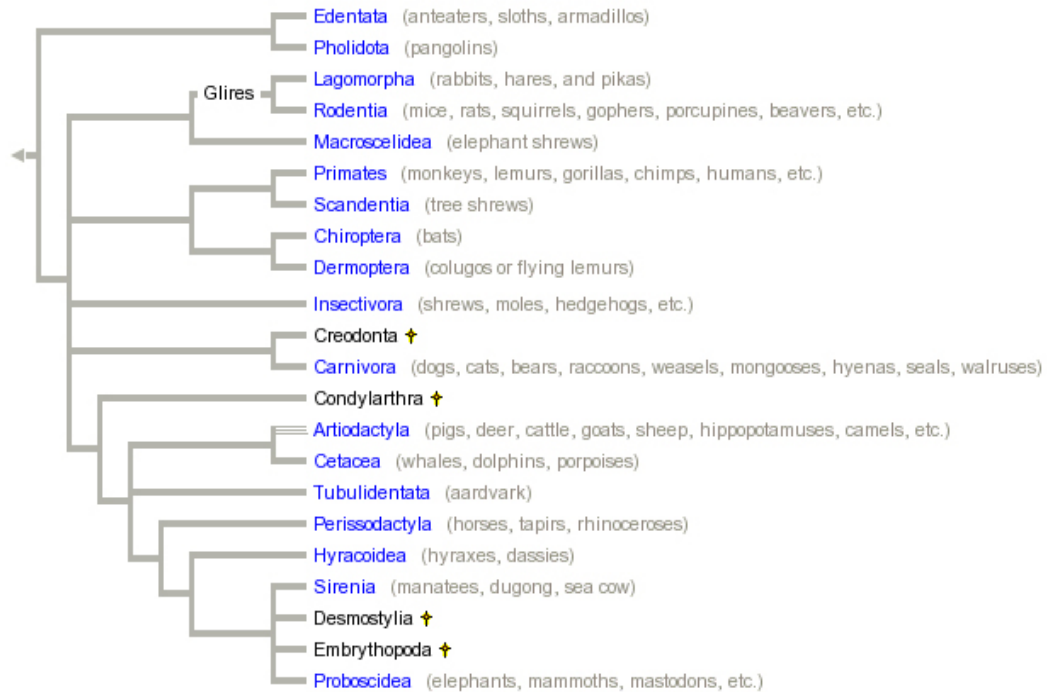
*"The affinities of all the beings of the same class have sometimes been represented by a great tree... As buds give rise by growth to fresh buds, and these if vigorous, branch out and overtop on all sides many a feebler branch, so by generation I believe it has been with the great Tree of Life, which fills with its dead and broken branches the crust of the earth, and covers the surface with its ever branching and beautiful ramifications."*

Charles Darwin, 1859

**Figure 5.9.** A section of the main page from the “Tree of Life Web Project” web site (<http://tolweb.org/tree/>, March 28<sup>th</sup>, 2006). The diagram in the middle of the page allows a browser to search for species within particular branches of the tree of life.

## Eutheria

Placental Mammals



**Figure 5.10.** A section of the ‘Eutheria’ (placental mammals) page (<http://tolweb.org/Eutheria/15997>, March 28<sup>th</sup>, 2006). Underneath the tree is a list of references of articles that were used in the creation of the tree. Clicking on one of the nodes brings up a new branch of more specific classifications, whereas clicking the arrow on the left brings up a new branch of broader classifications.

**Table 5.4** Analogical comparison while making an addition to a phylogenetic tree.

<b>Species to be added</b>	<b>Species in Phylogenetic Tree</b>
Species <sub>a</sub>	Species <sub>p</sub>
Sequence <sub>a</sub>	Sequences <sub>p</sub>
Similarity score (sequence <sub>a</sub> , sequences <sub>p</sub> )	Similarity score (sequences <sub>p</sub> , sequence <sub>a</sub> )
Genus <sub>p</sub> (species <sub>a</sub> )	Genus <sub>p</sub> (species <sub>p</sub> )

As we see in table 5.4, since the added sequence has a particular similarity score with the sequences from the species in the phylogenetic tree, the species that is to be added likely belongs to the same genus as the species in the phylogenetic tree.

A possible problem with this analogy is that when a new sequence is compared to the sequences from species already in a phylogenetic tree a new multiple sequence alignment needs to be performed. As a result, the distance scores that were previously calculated may change, thus possibly changing the tree itself. Consider the following sequences.

ABCDEF

ABCDEG

IBCDEH

A multiple sequence alignment performed on these three sequences would have the ‘ADCDE’ residues as conserved residues, meaning that any deviation would carry a high penalty. Therefore, an MSA would find the first two sequences more closely related to each other than to the third sequence. However, consider what happens when this fourth sequence is added:

JBCDEH

With this addition, the ‘A’ residue in the sequence becomes less likely to be conserved, and may thus change the configuration of the tree. As a result, the ‘known’ species classifications are actually somewhat undetermined, and can be classified differently with the addition of new species.

Many algorithms have been designed to account for this possibility. Also, because the branches of a phylogenetic tree are relatively well established, such a possibility would be relatively rare. For example, if a particular branch has about 10 species, the addition of a new species is unlikely to have a significant effect. The classification of the species in this branch is likely to be well supported by existing phylogenetic evidence, and thus a researcher would be more confident in saying that the configuration of this branch is correct.

It is apparent in the creation of phylogenetic trees using bioinformatics tools that analogical reasoning is accomplished in large part by the use of extended cognition. The use of computer algorithms such as ClustalW and MrBayes as well as web sites such as the Tree of Life web project have helped create modern phylogenetic trees, which allow researchers to make analogical inferences about species based on their related species. Just as with BLAST, this combination of computer use and analogical reasoning may be unique to bioinformatics research.

**5.11 Epistemic Appraisal of Phylogenetic Tree Creation.** The advantage of using extended cognition in the creation of phylogenetic trees is also apparent when one uses Goldman’s (1992) and Thagard’s (1997) appraisal standards for epistemic practices:

*Reliability:* Programs that are used to make phylogenetic analyses, such as ClustalW and MrBayes have proven to be quite reliable in generating scientifically acceptable results. These programs have become the favoured tools for creating phylogenetic trees over simply using fossil records or even comparing morphological/behavioural/sexual traits among species. The reason for this increased reliability has to do with the precise and theoretically acceptable algorithms used by these tools, especially the algorithms that generate multiple sequence alignments. These alignments are precisely designed to find conserved genetic sections in a number of sequences, and to calculate species divergence based on the sequence divergence from these conserved sections. These tools are not restrictive in their use, however, thus allowing researchers to create trees based on their preferred parameters (see figure 5.4).

*Power:* Due to the fact that these tools employ algorithms that are specifically designed to compare biological sequences, as well as create phylogenetic trees based on those comparisons, these algorithms generate many scientific results. All a researcher requires in order to construct phylogenetic trees are reliable genetic sequences from a set of species as well as algorithms in order to compare those sequences and construct trees from those comparisons. These methods are much more powerful than previous methods that required detailed information on the phenotypes of various species and/or fossil evidence from ancestral species.

*Fecundity:* As presented in figures 5.5 and 5.8, there is a plethora of results returned in the creation of phylogenetic trees using programs like ClustalW. Not only are the trees created, but results are given for how those particular trees were created by presenting the sequence comparisons using the multiple sequence alignments, as well as

the calculated distance metrics. An advantage of using these bioinformatics tools, as well, is that a researcher does not require access to limiting information, such as fossils records or direct observation of a species' phenotypic traits. Many researchers have access to many organisms' sequence information via specific databases like GenBank, and can thus create trees using any species they wish. Thus, the potential to create many more phylogenetic trees is much greater than before these bioinformatics tools were available.

*Speed:* Bioinformatics tools like ClustalW are relatively quick compared to previous methods, although these algorithms are not nearly as fast as BLAST. The only lengthy procedure is collecting the sequence information, and even this step requires only minutes of a researchers time. Once the sequence information is found, creating a tree like the one in figure 5.8 takes less than a minute. Searches with a larger set of sequence data would certainly take longer, yet not nearly as long as a non-bioinformatic approaches such as using fossil records or comparing phenotypic traits.

*Efficiency:* Just as with BLAST, the cost of creating phylogenetic trees using bioinformatics tools is minimal. All a researcher requires in order to create a phylogenetic tree is a desktop PC or laptop with any kind of Internet connection. In fact, it would be much more cost-efficient for a researcher to invest in computers that perform the analyses rather than attempt to find fossils, or even collect and analyze the phenotypic traits of the species in question.

*Explanatory Efficacy:* Programs that create phylogenetic trees, like the 'Tree of Life', help create coherent pictures of how all species are related to one another. These programs, I believe, also provide better explanations than previous methods of how these species are related to one another. Previous trees were created based on fossil and



phenotypic similarities, yet many people did not see how these similarities necessarily amounted to evolutionary relationships, and rightfully saw large gaps in species similarities. For example, birds are thought to have evolved from reptiles, yet many question this relationship due to the fact that there is minimal fossil or contemporary evidence of reptiles with ‘half-wings’. Also, any evolutionary explanation for the existence of these half-wings would be conjectural due to the fact that these species are extinct. If one instead compares genetic sequences, the relationships between species are much more straightforward, since it is easier to see how genetic sequences change than it is to visualize how phenotypic traits change. As explained earlier, a small change in a species’ genotype can cause very large changes in its phenotype, thus explaining how species that have many phenotypic differences can nonetheless have few genotypic changes. Thus, creating phylogenetic trees based on these genotypic changes helps better explain the evolutionary relationships among species.

**5.12 Summary.** In this chapter, I have summarized how bioinformatics research is used to create phylogenetic trees. This process uses extended cognition through computer use and analogical reasoning. The summary included a historical overview from Aristotle’s work on definitions, to Darwin’s work on evolution, to Hall’s review of the generation of trees using data from multiple sequence alignments. Current research into phylogenetic trees uses bioinformatics tools that create multiple sequence alignments of sequences from related species, and those alignments are then used to create either distance-based or character-based trees.

Although phylogenetic trees were previously generated without the use of computers, bioinformatics research has made the trees much more precise, and also much easier to generate. I showed how analogical reasoning is used in some aspects in the creation of phylogenetic trees, namely with the addition of new species to phylogenetic branches that already contained a number of species. It was also shown how analogical reasoning is performed through the use of computers. Lastly, I demonstrated how the creation of these trees using bioinformatics tools meet Goldman (1992) and Thagard's (1997) standards for the appraisal of an epistemic practice.

## Chapter 6

# DNA Microarrays Case Study

**6.1 Introduction.** As we have seen in the BLAST and phylogeny chapters, one of the major sources of information in bioinformatics is the large store of genomic data that is available. Along with the completed human genome sequence, there are completed sequences of many other organisms. Among organisms whose genomes are not completely sequenced, there is still a large set of sequenced DNA. The NCBI database contains over 100 billion bases from over 165,000 organisms, and the input of information is still growing exponentially (figure 4.1).

Despite this large set of genomic data, BLAST and phylogenetic comparisons are vastly limited in the type of analyses they are able to perform. Because BLAST and phylogenetic programs are only capable of comparing genomic and protein sequences, they can only show whether sequences are related, or whether a particular sequence is a candidate gene once other similar sequences have been identified as genes in a wet lab. If these were the only tools available to bioinformaticists, their discipline would still be very important. However, there is one more tool that has made bioinformatics indispensable, not only to biologists, but to medical researchers as well. That tool is the DNA microarray.

The microarray is a relatively new tool that helps answer one of the most fundamental questions in biology: What are the underlying mechanisms at work in a biological cell? Not only are microarrays a new tool for answering this question, they

have produced unprecedented amounts of information on the workings of biological cells. The use of microarrays is also called *high-throughput screening*, due to the large amounts of data that can be produced in a relatively short amount of time. However, the reliability of this tool is still very questionable.

The power of microarrays comes from the three previously described scientific methods: extended cognition through computer use, analogical reasoning, and representations of mechanisms. With respect to extended cognition, the requirement of computer processes and databases in both the generation and analysis of microarray data is so extensive that these computer components are indispensable. Analogical reasoning is not as apparent with this tool, but like BLAST and phylogenetic programs, microarray results come from comparing data that are unknown to data that are known. The types of operations that are possible with these tools include bio-molecular comparisons between species, organisms, tissues, and specific cells, along with many others.

The power of microarrays has allowed for biologists to generate theories that are much more complex than have ever been seen before. Generally, bioinformatics research has been used to generate *interactomes*, which are complex representations of mechanisms that show the interactions between numerous biological molecules, such as genes, proteins and other bio-molecules. These interactomes resemble massive networks with numerous nodes and their interconnecting causal relationships.

This chapter will look at the history and the underlying science behind microarrays, including some major discoveries made by using this technique. I will then look at how microarrays use extended cognition and a technique similar to analogical reasoning, and at how microarrays facilitate the creation of representations of complex

mechanisms. Lastly, an epistemic appraisal of microarray use will be done using Goldman (1992) and Thagard's (1997) standards.

## 6.2 History of Microarrays. A list of historical milestone is in Table 6.1.

**Table 6.1** Microarray Milestones (compiled from Southern, 2001, and Brewster et. al., 2004):

Year	Milestone
1988	Edwin Southern files UK patent applications for in situ synthesized, oligonucleotide microarrays.
1991	Stephen Fodor and colleagues publish photolithographic array fabrication method.
1992	Undeterred by NIH naysayers, Patrick Brown develops spotted arrays
1993	Affymax begets Affymetrix
1995	Mark Schena publishes first use of microarrays for gene expression analysis, Edwin Southern founds Oxford Gene Technologies
1996	First human gene expression microarray study published, Affymetrix releases its first catalog GeneChip microarray, for HIV, in April.
1997	Stanford researchers publish the first whole-genome microarray study, of yeast
1998	Brown's lab develops CLUSTER, a statistical tool for microarray data analysis; red and green "thermal plots" start popping up everywhere
1999	Todd Golub and colleagues use microarrays to classify cancers, sparking widespread interest in clinical applications
2000	Affymetrix spins off Perlegen, to sequence multiple human genomes and identify genetic variation using arrays
2001	The Microarray Gene Expression Data Society develops MIAME standard for the collection and reporting of microarray data
2003	Joseph DeRisi uses a microarray to identify the SARS virus, Affymetrix, Applied Biosystems, and Agilent Technologies individually array human genome on a single chip
2004	Roche releases Amplichip CYP450, the first FDA-approved microarray for diagnostic purposes

The initial concept for microarrays was developed in the late eighties separately by Edwin Southern and Stephen Fodor. Southern is generally known for developing the Southern blot test in 1975 (Southern, 1975), which is a DNA analysis tool that is

commonly used to identify genetic markers, and are used in paternity tests and criminal investigations. In his initial work on microarrays, Southern developed a method whereby inkjet printing was used to print the four amino acid nucleotide bases into oligonucleotide sequences on glass slides, similar to how inkjet printers use coloured inks. Fodor was working on a similar project where he used photolithography, which was a technique that etches the tiny features on semiconductor chips. Fodor and Southern also started companies in order to fund their projects: Fodor started Affymax, which later evolved into the currently popular microarray company, Affymetrix, while Southern founded Oxford Gene Technology.

Both Southern and Fodor recognized the potential use of their ideas, but the technology required was still not available. One of the basic requirements for a microarray to work is to have a plate with an oligonucleotide that can bind with a complementing oligonucleotide. By 1991, Fodor was able to build arrays of peptides and dinucleotides, but they contained sequences just eight nucleotides in length, less than a third of the standard oligonucleotides Affymetrix uses today. The eight-nucleotide long peptides are too short to bind with complementing molecules. However, the process was a step in eventually developing a functioning microarray.

In 1992, Patrick Brown, then at Stanford University, had a different strategy. Instead of synthesizing the molecules onto surfaces, which was the strategy being developed by Southern and Fodor, he envisioned lining up already synthesized oligonucleotides on the surfaces of slides. Although his idea was not supported by the grant agencies (his 1992 NIH grant proposal was summarily rejected), Brown continued to work on his idea. He built robots that would deposit small amounts of DNA on a glass

surface. The first attempts were horrible failures due to his lack of resources, but small improvements were always being made.

Mark Schena, a graduate student at Stanford in 1992, saw potential in Brown's spotted array for gene expression studies. At that time, gene expression studies were very limited, where scientists were only able to study the expression of one to a few genes at a time, and these studies involved very sensitive work in wet labs (gene expression will be described in greater detail in the next section). Brown and Schena began to work together to match the technology with a particular biological problem. They headed some groundbreaking experiments which led to the landmark 1995 *Science* paper that used the word "microarray" for the first time (Schena et. al., 1995). In 1997, the Brown lab published their first paper on the expression of a whole genome (Lashkari et. al., 1997).

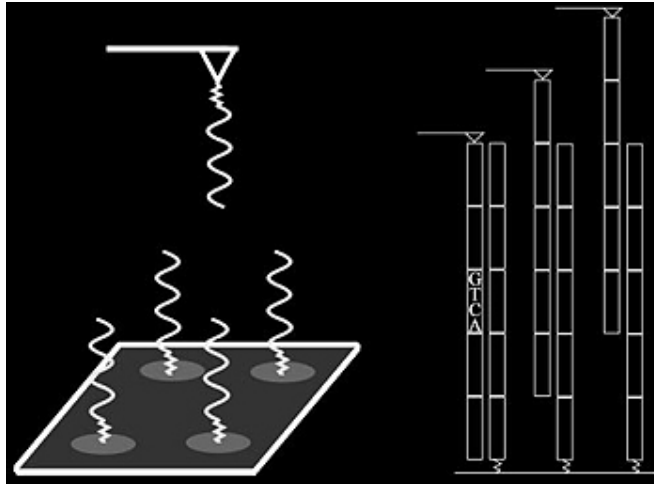
Since then, microarrays have been further refined to give more reliable results, and, more interestingly, have been used to solve many different biological problems. Along with biological applications, there has been much more interest into how microarrays can be used for medical discoveries. The first indication of this applicability was presented in Golub et. al.'s paper (1999) when they used microarrays to classify cancers. Advances in cancer treatment are currently very much dependent upon proper cancer classification. One of the most recent methods of classifying cancers is to determine the different effects it has on the human body. Since microarrays are able to measure changes in gene expression, then they are also a good measure for measuring the changes that occur within a cell. Golub et. al. (1999) therefore used microarrays to analyze molecules from cancer infected cells from numerous patients and thus were able to accurately classify the various types of cancer. The technique has become very

popular. Potti et. al. (2006) have developed a technique where microarrays help in determining the most effective choice of chemotherapy.

Another medical application is the recognition of unknown viruses. One of the most famous recent cases is that of SARS, which was identified by Joseph DeRisi's lab. (Wang et. al., 2003) using microarrays. Their method was to create DNA microarrays that contained oligonucleotides about 70 nucleotides long, and these oligonucleotides were derived from every fully sequenced reference viral genome in GenBank at the time. By separating the SARS RNA and allowing it to bind to the microarray, the researchers were able to show that the SARS virus was a previously unknown virus as well as identifying its closest known coronaviral relative. By knowing its genetic relative, medical researchers can proceed with clinical trials of the drug used to treat its closest relative on patients infected with SARS. These last two applications will be described in greater detail after we look at exactly how microarrays work.

**6.3 How Microarrays Work.** As was discussed earlier, the theory behind microarrays is relatively simple. The difficult part is coming up with the proper technology to realize the theory. The most general way to describe microarrays is that they are molecules on a glass plate. To give a mental picture of the apparatus, the molecules 'stand up' on the plate. The purpose of this 'standing' is to allow any molecules that come into contact with the plate to bind with plated molecules. As such, microarrays are a lot like Velcro. See figure 6.1.



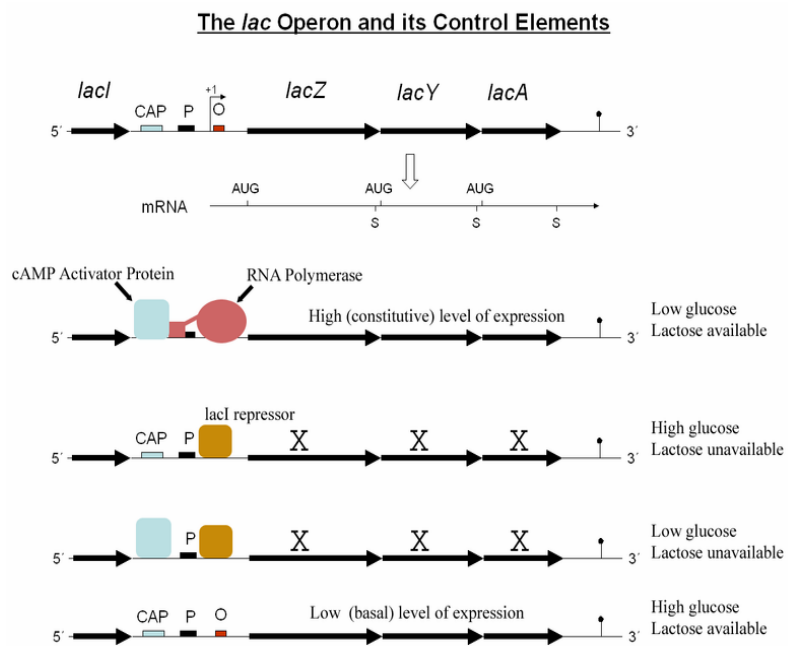


**Figure 6.1.** A close-up of a side view of a microarray surface. Oligonucleotides ‘stand’ from the surface, thus making it possible for them to bind to any incoming molecules. In this diagram, the plate has three kinds of oligonucleotides. A typical microarray contains thousands of clones of each oligonucleotide (Sattin et. al., 2004).

The molecular structure of the oligonucleotides is known for most types of experiments that use microarrays. The types of oligonucleotides that are plated depend upon what the purpose of the experiment will be; however, most experiments attempt to find differences in gene expression among cell lines. A little aside is required at this point to explain gene expression and its importance.

Along with an organism’s genome, there are genetic ‘switches’ that control which genes are transcribing proteins at a particular moment. One of the first studies to show the workings of gene expressions was that of Monod and Jacob in 1959 (Baldi & Hatfield, 2002). They worked on what they termed the *lac operon*, which was a group of genes that was responsible for the conversion of lactose into energy in *E. coli*. They found that when lactose was presented to the bacteria, it attached to an area called the *promoter* on the *E.*

*coli* DNA. Once the lactose was bound, this signaled the *operator* on the operon to allow for the production of the gene that transcribes lactase, which is the protein needed to break down lactose into its constituent galactose and glucose molecules, which can then be converted into energy. If no lactose was present, then the production of that gene was inhibited. This inhibition occurs since natural selection normally favours species that are efficient, and in this case, favours species that do not produce molecules that they do not need. If, on the other hand, glucose were presented to the bacteria, then the glucose would bind to the *repressor* on the operon, which would not allow any lactase to be produced. Even if there were any lactose, the lactase would only be minimally produced, since the bacteria would not need to break down the lactose if there already is glucose present (see figure 6.2).



**Figure 6. 2.** The *lac* operon, and the production of lactase under different cellular conditions (diagram permission granted under the Creative Commons Attribution License).

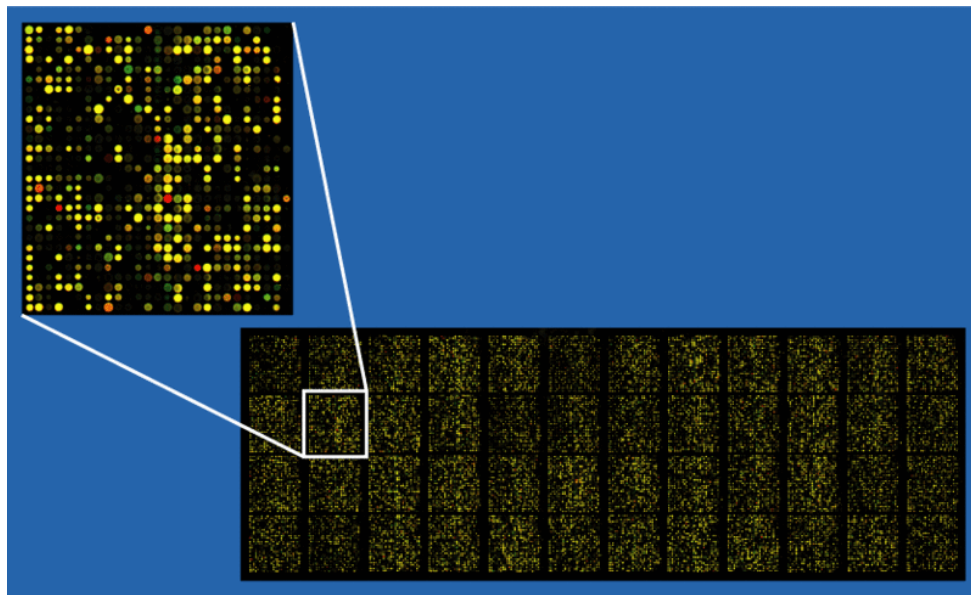
Generally speaking, gene regulation is the method by which a cell turns on or off particular genes. Without gene regulation, genes would constantly remain in either an ‘on’ or ‘off’ state. Gene regulation is important for many different cellular tasks: cell development, responses to environmental conditions, responses to diseases, and in multi-cellular organisms, gene regulation is also the means cells use to differentiate from each other and develop into different organs. For example, human liver cells and human brain cells have the same genome within each cell, but they are differentiated due to which genes are turned ‘on’ in those cells, and at which times.

Understanding gene regulation now helps us understand why microarrays are important. Microarray researchers begin with plates composed of known nucleotides, usually the entire genome of an organism, or whatever genes they have sequenced from that organism so far. They then ‘wash’ the plates with radioactively labelled mRNA that are produced within particular cells, and/or cells that are found in particular conditions. The mRNA found within the cells binds with the oligonucleotides on the plate, and these bound nucleotides would emit the radioactivity of the mRNA. Using robotic and computerized detection techniques, researchers can identify which nucleotides have bound-mRNA, and the intensity of the radioactive signal indicates how much mRNA is produced in the cell. Thus, any bound nucleotide is a gene that is expressed in the cell. Microarrays allow researchers to measure the simultaneous mRNA expression levels of the thousands of genes represented on an array as well as the intensity of that expression (see figure 6.3).

This simple experiment, however, does not yet yield the information researchers typically look for when using microarrays. Let’s say a researcher washes a microarray

with the mRNA products from one cell. The researcher would discover which genes are being expressed in that cell, but they would not necessarily know which of those expressed genes make that cell unique from other cells. In order to discover the cell's uniqueness, the researcher would require another cell line to compare with the original cell line (much like a control). There are many types of comparisons that are possible:

1. Normal cell lines with diseased cell lines.
2. Normal cell lines with mutated cell lines.
3. Cell lines from tissue *A* with cell lines from tissue *B*
4. Cell lines at developmental stage *C* with cell lines at developmental stage *D*.
5. Cell lines from similar organs from different species.



**Figure 6.3.** A microarray with radioactive mRNA washed over the surface. Each spot represents a different gene, and the particular colour of each spot represents (not shown here) the relative expression levels of the gene in 2 different cell types (diagram is in the public domain).

These comparisons are not necessarily limited to examining two cell lines, but can involve many cell lines. For example, in comparing normal cells with diseased cells, a researcher may want to test the diseased cells at different times.

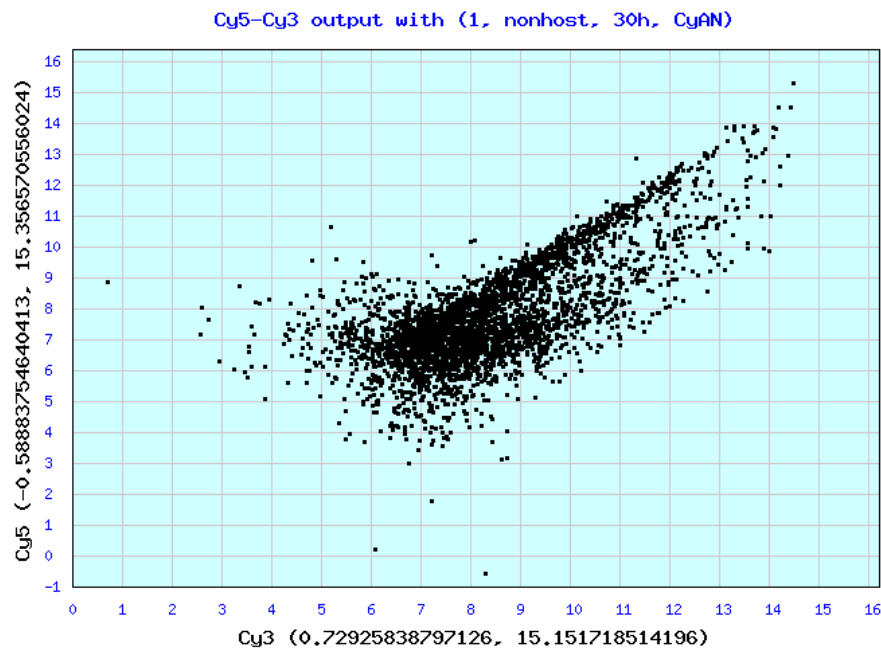
There are many different methods using microarrays to perform these comparisons, but I will only describe the method that I believe is the simplest in order to keep the description of microarrays short. Using different radioactive dyes on the mRNA from each cell line, the researchers can go about comparing levels of expression on one DNA microarray (Pevsner, 2003). The technique uses principles from elementary art class, which tells us how a resulting color may be dependant upon a combination of simpler colours. Typically, a red dye (Cy5 radioactive dye) is used on one cell line and a green dye (Cy3 radioactive dye) is used on the other cell line. When a gene from one line is expressed more than in another line, the resulting color on the microarray will be closer to the initial dye color used on that cell line. If the expression levels are the same, then the spot is either black, meaning there was no expression in either line, or yellow, which is the combination colour of the two dyes that are used in this case.

The best way to describe how the process works is by presenting a specific case. Cell line A is the test group, where a virus is introduced to the cell line. The mRNA produced by the cell line is labeled with green dye. Cell line B is the control group, meaning that it is the same as line A except that it does not have the viral infection. The mRNA from line B is labeled with red dye. A microarray is prepared where each gene of the cell's genome is a spot on the chip. The chip is 'washed' with the labeled mRNA from each cell line, allowing the mRNA to bind with the DNA already on the chip. The resulting chip can be something very similar to what is presented in figure 6.3 except

with colour, where some spots are black, some green, some red, and some yellow. Of course, the colours of each spot are not as easily divisible into this four colour distinction. Genes are expressed at different intensities, and are not simply either 'on' or 'off'. If one were to look closely at these washed chips, one notices that some spots are intensely green or intensely red, some mildly green, some mildly red, some greenish-yellow, some reddish-yellow, and so on. Thus, there is more of a spectrum of colours. What needs to be done next, then, is to have an accurate analysis of the colours that are presented on the chip.

**6.4 Image and Data Analysis.** Image analysis measures the amounts by which the expression of particular genes change between the cell lines that are under comparison. To reiterate, the purpose of microarrays is not necessarily to measure the amount of gene expression in a cell line, but the difference in expression among cell lines. A spot that is intensely green means that the gene is heavily expressed in cell line A but not expressed in cell line B. A spot that is intensely red means that the gene is heavily expressed in cell line B but not in cell line A. A black or yellow spot means that the gene is expressed by equal amounts in both lines. The intensity of each spot is analyzed using sophisticated lasers that are able to detect the intensity of each dye (Cy5 and Cy3) rather than the combined colour. For each spot, a value is given for the intensity of the green dye and a separate value for the intensity of the red dye. These intensities are plotted on a graph, and this allows researchers to determine which of these intensities fall outside the 'normal' range. Figure 6.4 shows such a graph. The red channel values are plotted on the y axis, which is the expression level of cell line B (non-infected) and the green channel

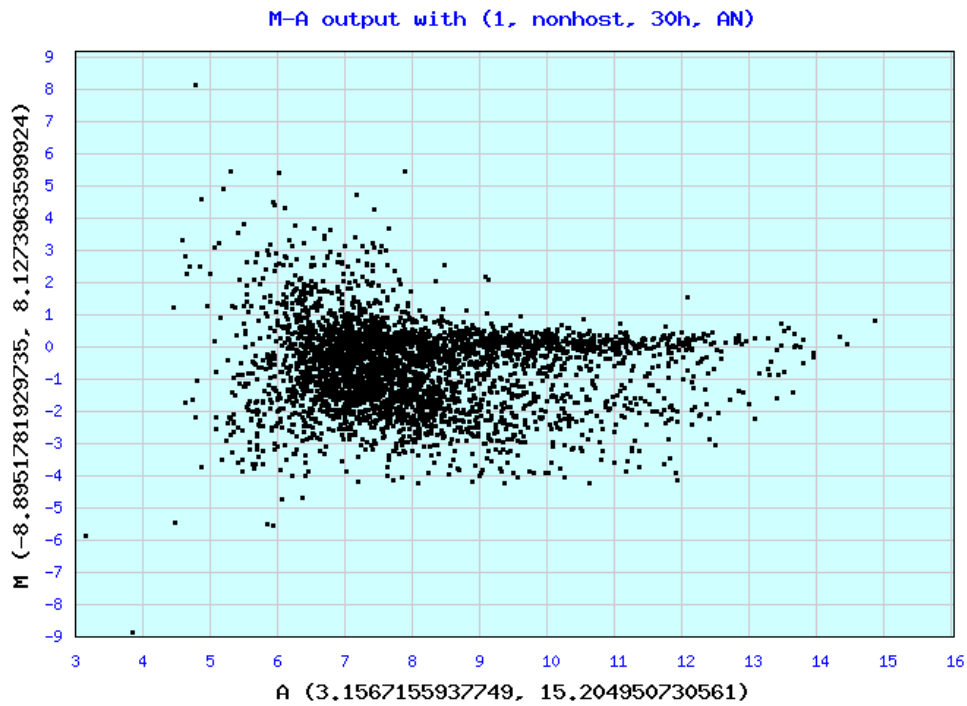
values are plotted on the x axis, which is the expression level of cell line A (infected). Genes that are expressed equally in both lines have the same or relatively close x and y values, whereas those that are expressed differently have different values. Expression levels that are closer to the 0,0 value mean that the gene is hardly expressed in both cell lines, whereas expression levels that have both relatively high x and y values mean that the gene is highly expressed in both cell lines. Normally, there are many more genes plotted near the 0,0 mark, meaning that most genes are not expressed at any given times (not shown in figure 6.4 since some transformations to the graphs were already made). Genes that have a lower y value than x value mean that there is a higher expression of that gene in the infected line. Genes that have a higher y value than x value mean that there is a lower expression of that gene in the infected line.



**Figure 6.4.** Graph comparing the expression levels of genes in different cell lines (figure from Efron et. al, 2001).

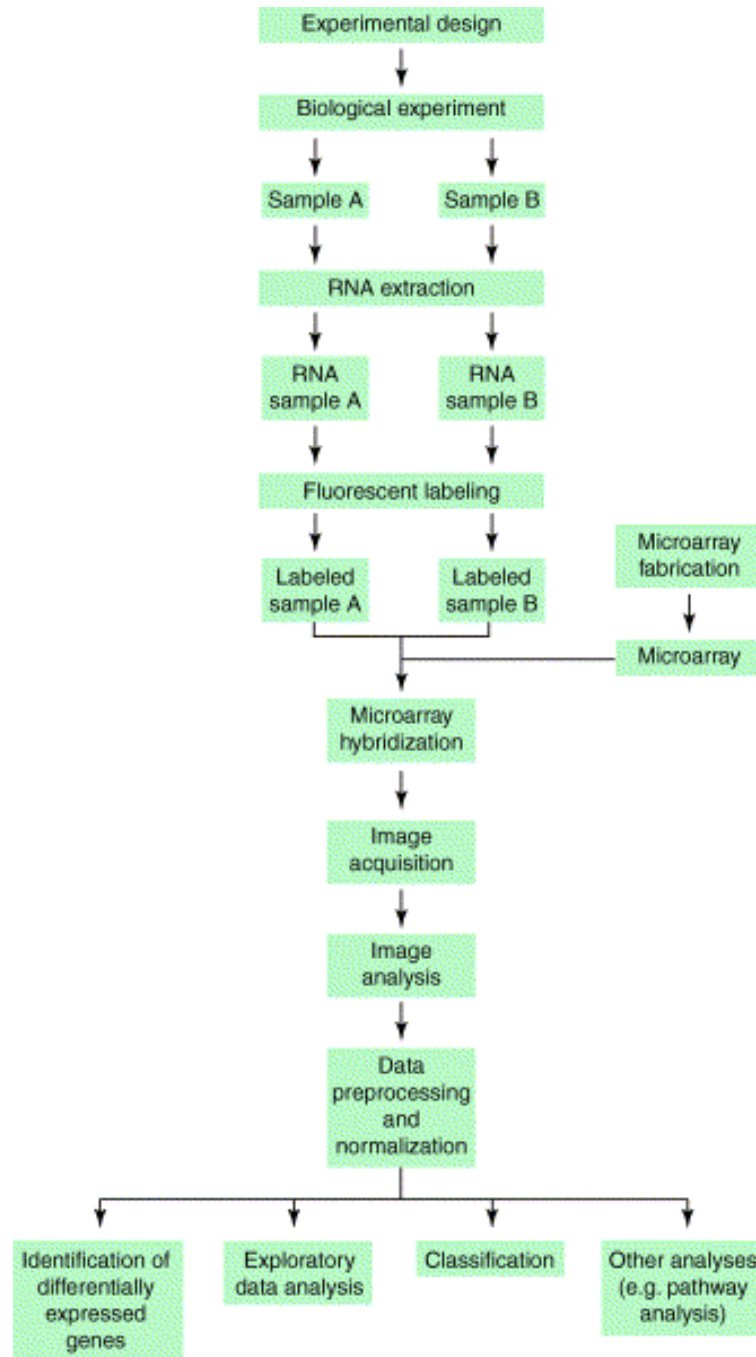
The next step is to use inferential statistical analyses tools to determine which genes are expressed significantly differently. There are many different types of statistical tools that are available to a bioinformaticist. Many begin by having axes converted from linear scales to logarithmic scales, so that the data set is a little more symmetrical along the x,y axis, and not clustered around the 0,0 value. Next, the data can be transformed in order to plot the geometric mean intensity versus the log of the gene expression value ratios, thus making it easier to point out which genes are expressed differently (figure 6.5). Many other transformations are possible in order to better account for errors that may occur in creating microarray analyses. For example, one notices from figure 6.4 that expression levels vary more in the areas of high expression than low expression. This may be due to problems in the mRNA collection as well as problems with the problem of dye absorption. Thus, specific statistical tools are used to reduce the possible errors caused by among genes that have high expression levels.





**Figure 6.5.** Normalized data of changes in levels of expression. The changes in expression levels are not measured by their deviation from the x, y axis, but from the log of the gene expression value ratios (figure from Efron et. al., 2001). Notice that it is easier to determine significant expression differences in this graph compared to the graph in figure 6.4.

Diagrams such as these give bioinformaticists information on the effect of particular conditions on particular cell lines, as in our case, the reaction of a cell to a viral infection. The whole process, from creating a microarray research project to the analysis of the data, is presented in figure 6.6.



**Figure 6.6.** A schematic representation of the microarray research study (figure from Leung & Cavalieri, 2003)

**6.5 Creation of Interactomes.** Microarrays are invaluable for the results I have described so far, yet there are even more valuable results that can be produced. As was described earlier in the chapter, microarrays are also credited with generating very complex interactomes. Interactomes are complex mechanisms that show the interactions among biological products. Although the interactomes generated by microarrays are not very precise, these interactomes are still a step in the direction of eventually discovering the complete inner workings of cells.

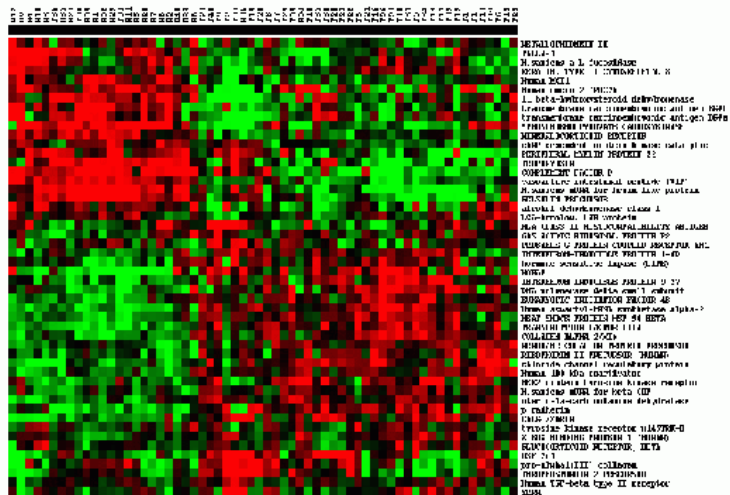
Previous methods at creating interactomes were complicated wet lab procedures that were very costly as well as time-consuming. The wet lab procedure to discover the components and their interactions is to devise methods to detect each possible component and discover what occurs when each is removed from the reaction. Although the process is very precise, it is both very costly and time consuming.

The advantage of the microarray is that researchers are finally able to see all at once the components that are important in particular situations. Using the example above, the mRNA products that are likely to be involved in responding to the viral infection are the ones that are up-regulated or down-regulated. Other genes may be involved, included those whose expression does not change, yet this is unlikely: An underlying assumption that seems to be at play in most biological research is that cells react to particular situations through changes in their gene expression. The RNA and protein products that are produced by cells take milliseconds to produce, but only last for a few minutes. Therefore, instead of producing RNA or proteins that continually break down without being used, the cell can easily react to situations as they present themselves, thus maximizing its energy and resources.

With the components of the mechanism at hand, all that is missing is how these components interact. Another assumption comes into play at this point: If two products are involved in the same mechanism, then their modes of regulation are likely to be somewhat similar. The graphs above give indications of the extent of these modes of regulation. For example, figure 6.5 shows that some products are up regulated by a certain amount, some down-regulated, and some staying the same.

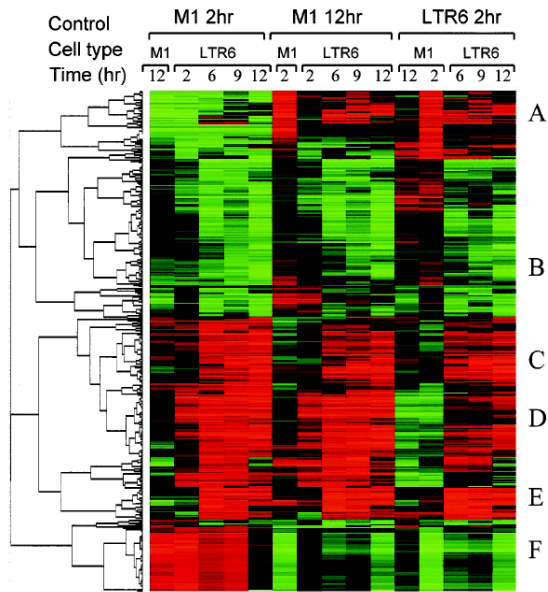
Computer programs have been designed in order to detect products that share modes of regulation. After detecting these products, they then group the ones that share regulation patterns. For example, if genes 1, 2 and 3 are up-regulated and gene 4 is down-regulated, then 1, 2 and 3 are put into one group and 4 is put into another. Lastly, there are programs that construct the possible interactomes that might exist based on these genes that are grouped together, creating interactions between those that share the closest regulation patterns and then creating possible interactions with those that have less similar regulation patterns. We will look at each of these steps in detail.

First, the detection software detects the expression pattern of each gene. Arrays are created that can represent the regulation pattern of each gene under each presented condition. These arrays can almost be as confusing as the microarrays themselves, as shows in the example below (figure 6.7):



**Figure 6.7.** Representation of the expression levels of a set of genes (rows) in different cell lines (columns). (From <http://dir.niehs.nih.gov/microarray/datamining/>, permission for use granted by Dr. Leping Li)

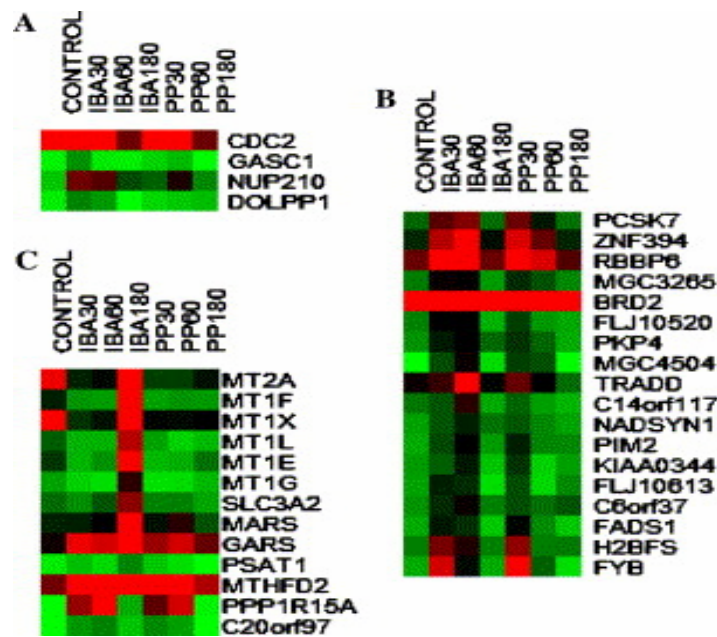
Although the figure above looks like a microarray, it is not. The figure above is a computer representation of the expression levels of particular genes, which are listed in the rows, under different conditions, or cell lines, which are listed in the columns. The genes that are up-regulated under particular conditions are represented by green dots that vary in intensity and the genes that are down-regulated under particular conditions are represented by red dots that vary in intensity. The representation, as it is presented here, does not clearly show which genes can be clustered together for having similar expression patterns. Computer algorithms have been created that can easily find the patterns based on the levels of expression. The clustering of the genes can be represented using the same type of graph seen below (figure 6.8).



**Figure 6.8.** Cluster analysis of genes that share regulation patterns. Those whose patterns are most similar are clustered first, and clustering continues to include all genes. (from [http://www.weizmann.ac.il/home/ligivol/apoptosis\\_project/apoptotic\\_pathways.html](http://www.weizmann.ac.il/home/ligivol/apoptosis_project/apoptotic_pathways.html), permission for use granted by Dr. David Givol).

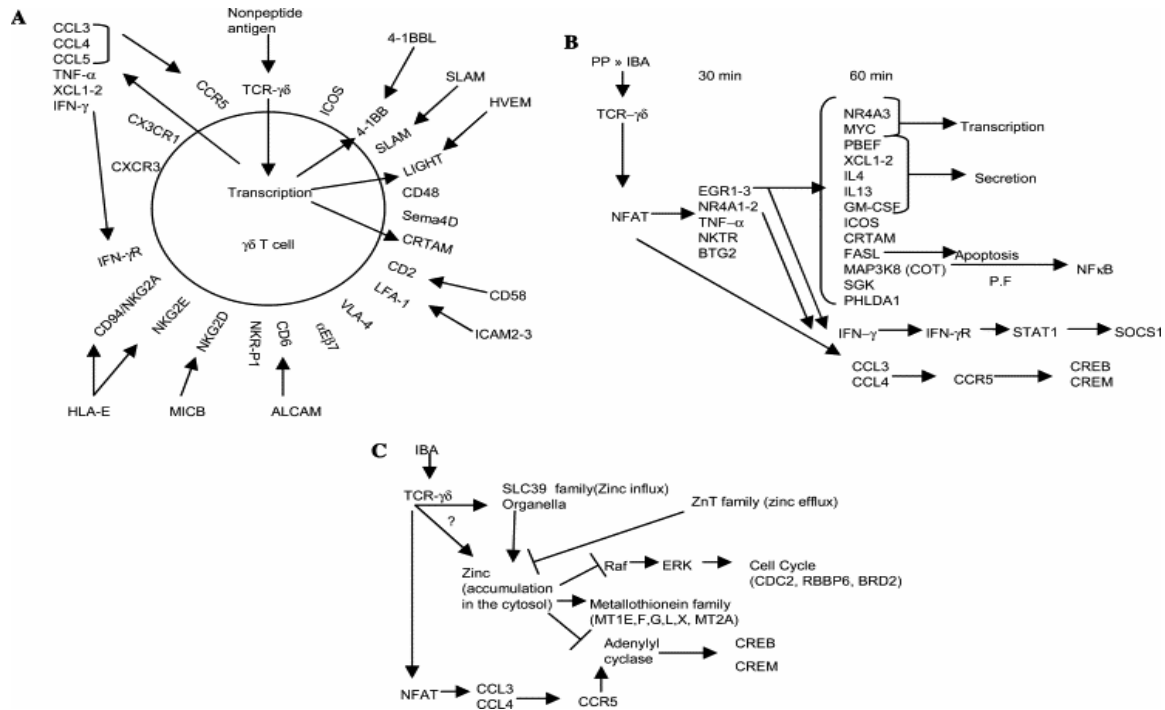
Notice in the above diagram that the genes that are closely clustered are the ones that have very similar expression patterns. Once the genes are clustered, then researchers can begin to construct interactomes based upon the patterns of clustering. The figure above shows three major clusters, thus pointing to three possibly separate mechanisms that are affected in this experiment. Within the major clusters, the genes that are most closely clustered are likely to be in a direct causal relationship than those that are more loosely clustered. Based on this distinction between close and loose clustering, researchers can construct elaborate interactomes that are likely to closely estimate the actual mechanisms that exist within cells.

A search through the on-line article search engine “Biological Sciences” with the terms “microarray” and “mechanism” alone return 361 hits, and with “microarray” and “pathway” return 441 hits, and it is likely that there are more papers where microarrays were used to find cellular mechanisms or pathways without those latter two terms being explicitly used in the paper. This shows that using microarrays to find these interactomes is a very popular technique. One example of a mechanism that was derived using microarrays is from Yamashita et. al (2005), which I briefly presented in the introduction. These researchers were testing the effect of specific non-peptide antigens on  $\gamma\delta$  T cells. They tested the reaction of these cells to the antigens at different times and were able to generate representations of gene expression such as the following (figure 6.9):



**Figure 6.9.** Representation of the expression of particular genes in T cells after being exposed to nonpeptide antigens (figure from Yamashita et. al., 2005).

Using expression data such as these, and clustering the genes according to the expression patterns, Yamashita et. al. hypothesized mechanisms such as the following (figure 6.10):



**Figure 6.10.** Mechanism generated from analyzing expression patterns (figure from Yamashita et. al., 2005).

Thus, from a biological perspective, microarrays are very useful in generating complex biological mechanisms, which are important theoretical tools for biologists.

Like the other bioinformatics tools that were presented in preceding chapters, online tools are available to generate new results from inputted microarray data. One of the most popular web tools is Gene Expression Omnibus (GEO) (Edgar et. al., 2002), which is, again, found on the NCBI web site at [www.ncbi.nlm.nih.gov/geo](http://www.ncbi.nlm.nih.gov/geo). The main search page of the program is found below. Researchers can deposit their microarray data



into the GEO database. For example, a researcher can deposit data from a microarray representing a human adult blood cell in ‘normal’ conditions, and another researcher can deposit data from a microarray representing a human adult hemophiliac blood cell. With this deposited DNA, a researcher can use the web tool to find the gene expression differences between the two microarrays. Other comparisons can be made as was described above, such as clustering genes with similar gene expression patterns and developing mechanisms from those clusters. Thus, once again, a researcher can generate biological results *in silico*.

Figures are presented below showing the microarray data publicly available. The two figures below show the main query page and returned pages after a query is submitted. The following examples have returned pages for comparing ‘normal’ adult human blood cells with those afflicted with a particular type of hemophilia (figures 6.11, 12).

NCBI

GE **Gene Expression Omnibus**

HOME SEARCH SITE MAP Handout NAR 2005 Paper NAR 2002 Paper FAQ MIAME Email GEO

NCBI > GEO Not logged in | Login

**Gene Expression Omnibus:** a gene expression/molecular abundance repository supporting [MIAME compliant](#) data submissions, and a curated, online resource for gene expression data browsing, query and retrieval.

**GEO navigation**

**QUERY**

- DataSets
- Gene profiles
- GEO accession
- GEO BLAST

**BROWSE**

- DataSets 
  - Platforms
- GEO accessions 
  - Samples
  - Series

**SUBMIT**

- Direct deposit / update
- Web deposit / update

**Public data**

GPL Platforms	2027
GSM Samples	69732
GSE Series	3042
<i>Total</i>	<b>74801</b>

**Site contents**

**Documentation**

- Overview | FAQ
- Web deposit guide
- Batch deposit guide
- Linking & citing
- Journal citations
- DataSet clusters
- GEO announce list
- Data disclaimer
- GEO staff

**Query & Browse**

- Repository browser
- Submitter contacts
- SAGEmap
- FTP site
- GEO Profiles
- GEO DataSets

**Deposit & Update**

- Direct deposit
- Web deposit
- New account

GEO help: Mouse over screen elements for information

**Get GEO accession**  Scope:  Format:  Amount:

**Depositors only** User :  Password :   [Recover a password](#)

**Figure 6.11.** The main page of GEO, found at [www.ncbi.nlm.nih.gov/geo](http://www.ncbi.nlm.nih.gov/geo) from Feb 17th, 2006. The researcher starts by inputting the data set that they are looking for through inputting its ‘DataSets’, ‘Gene Profile’ or ‘GEO accession’ identification. In the following figures, I have inputted the GEO Accession of GDS761, which represents a particular hemophilia comparison.

NCBI *DataSet Record* GEO Gene Expression Omnibus

HOME SEARCH SITE MAP NAR 2005 Paper NAR 2002 Paper FAQ MIAME Email GEO

NCBI > GEO > GDS

### GDS Summary

<b>Accession:</b>	GDS761 <a href="#">View Expression (GEO profiles)</a>		
<b>Title:</b>	Essential thrombocythemia expression profiling		
<b>DataSet type:</b>	gene expression array-based (RNA / in situ oligonucleotide)		
<b>Summary:</b>	Expression profiling of malignant megakaryocytes (MK) from 6 donors with essential thrombocythemia (ET). MK derived from bone marrow CD34+ hematopoietic progenitor cells. Progenitors induced to differentiate into MK by treatment with 100 ng/ml thrombopoietin. Apoptotic pathway is impaired in ET.		
<b>Platform:</b>	<a href="#">GPL96: Affymetrix GeneChip Human Genome U133 Array Set HG-U133A</a>		
<b>Sample organism:</b>	Homo sapiens	<b>Platform organism:</b>	Homo sapiens
<b>Feature count:</b>	22283	<b>Value type:</b>	count
<b>Series:</b>	<a href="#">GSE997</a>	<b>PubMed ID:</b>	<a href="#">15271793</a>
<b>Series published:</b>	01/26/2004	<b>Last GDS update:</b>	11/10/2004

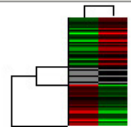
### Subset and Sample Info

**Sample selection**

check all    uncheck all    toggle

**Data**

download    analysis



2 assigned subsets			Two-tailed t-test (A vs B)	
Samples	Type	Description	A	B
<input checked="" type="checkbox"/> (1)	disease state	normal	<input type="checkbox"/> 0.050 significance level	<input type="checkbox"/>
<input checked="" type="checkbox"/> (1)	disease state	malignant	<input type="checkbox"/> ↔	<input type="checkbox"/>
<input checked="" type="checkbox"/> GDS761 only <input checked="" type="checkbox"/> ranks <input checked="" type="checkbox"/> values   subset effects			<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>

2 samples, order: none

[GSM15648](#) : Normal Megakaryocytes  
src1: Bone marrow, TPO treated, CD34 positive cells (6 donors pool)

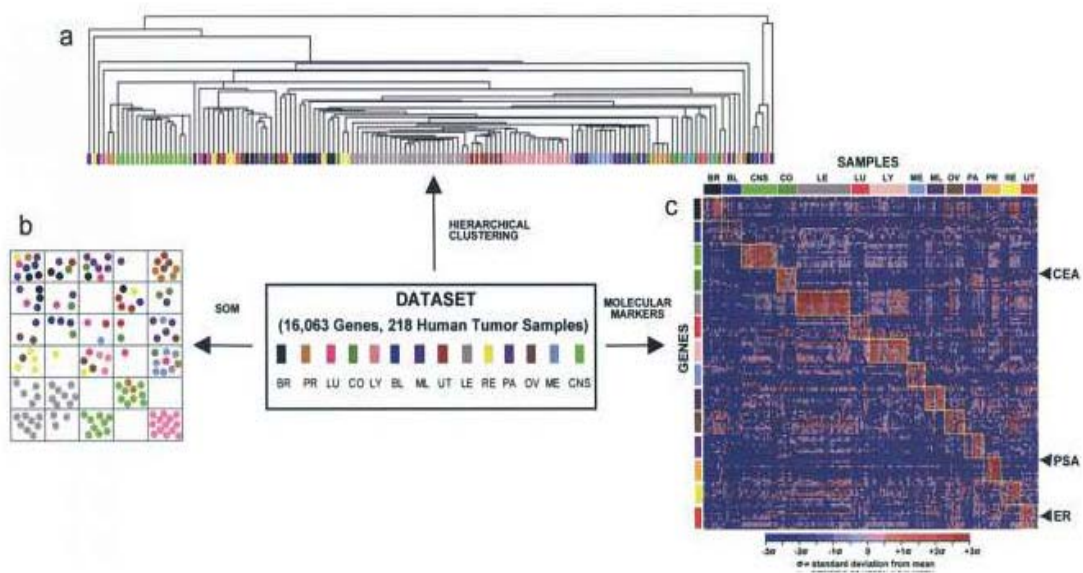
[GSM15650](#) : Essential Thrombocythemia (ET) Megakaryocytes  
src1: Bone marrow, TPO treated, Cd34 positive cells (6 ET patients pool)

**Figure 6.12.** Page showing data from GEO accession # GDS761. It has information on how the data was collected, what test subjects were used, and allows one to perform a variety of tests on the data sets, including t-tests, which give results on which genes are expressed significantly differently.

Therefore, microarrays are used to significantly advance biological knowledge by examining gene expression and developing representations of mechanisms using the measured expression levels. However, their use is not only restricted to biological research, as will be demonstrated in the next section.

**6.6 Medical Uses of Microarrays.** As mentioned in the introduction of this chapter, there have been some very important medical developments using microarrays, and the use of microarrays may soon become an indispensable tool for medical researchers. One recent major discovery using this tool is the creation of precise classifications of cancers. This precise classification can further help develop drugs that are specific to particular cancers. Another indication that microarrays may become essential tools is their ability to quickly identify new viruses. They help researchers find effective treatments to those new viruses quicker than previous methods.

Golub et. al. (2001) created a gene expression database with the expression profiles of 14 common human cancer classes. This was done by placing the mRNA from the different cancers on microarrays containing genes that are normally implicated in playing a role in cancer. The expression levels of these mRNA products differ among the different cancers. By finding the molecular bases of cancers, medical researchers will be able to better diagnose and treat the various classes of cancer. Figure 6.13 shows the results from their microarray analysis:



**Figure 6.13.** Expression of different cancer classes. The clustering diagram at the top helps in showing which genes are expressed most differently in each cancer case (figure from Golub et. al, 2001).

The second major medical discovery was the annotation of SARS by DeRisi et al. in 2003. DeRisi et. al. expanded the microarray technique to not only look at the expression of genes within a genome under different conditions, but to quickly compare multiple genomes. They put a number of oligonucleotides from all previously sequenced viral genomes on a plate, totaling about 10,000 different oligonucleotides. They then split the RNA from the unknown SARS genome and washed it over the surface of the microarray. It was thought likely that the SARS genome is similar to already sequenced genomes, and thus would bind to the oligonucleotides to which it shares its sequence. Using this technique, the researchers were able to show that SARS was very similar to the class of IBV coronaviruses. Subsequent studies that sequenced the virus were able to confirm these findings. The advantage of this technique was that the researchers were

able to discover, within 24 hours, that the virus was a novel organism and to what viruses it was related. Other techniques to generate the same results would have taken at least a week of gene amplification, sequencing, and comparisons, the last being accomplished through another bioinformatics tool, BLAST. Although this may not seem significant to a biological researcher, it is an important development in the medical field, one which could potentially save many lives during viral outbreaks.

This type of research is fundamentally different from the traditional use of microarrays which compared arrays of sequences to find differences in gene expression. Instead of finding differences among molecules, this research is looking for similarities among viral genomes. It is interesting to note that although microarrays were only developed recently, they are being used for fundamentally different types of research.

**6.7 Extended Cognition in Microarray Studies.** Almost every stage in microarray research requires information from computer databases or has information manipulated by computers. Just like the other bioinformatics tools studied, computers are indispensable for carrying out the operations required in research using microarrays.

All microarray research begins with the creation of the microarray itself. Normally, microarrays contain cDNA spots that represent most genes within the genome of a particular organism. The microarrays may also contain a number of genomes: The SARS example presented in the last section required a microarray with a number of oligonucleotides from a number of known viruses, totaling 10,000 different spots. This first step requires the following computer databases: Databases containing the gene

sequences of known genomes and databases created during the experiments that specify the location of each cDNA representing the gene on the microarray.

Once these databases are created, the following computer/robotic manipulations are carried out:

- 1) Placing a spot of each cDNA on a microarray chip.

No single scientist could memorize the gene sequence information found in the computer databases, and neither can they memorize the location of each cDNA placed on the microarray. The robotic placement of the cDNA on the microarray chips is an extremely complex and fragile process, one that, despite many technological developments, is still a source of problems in the creation of microarrays. However, these problems are often taken into account in further statistical analyses performed on microarray results in order to minimize them.

The next step in microarray research involves washing mRNA products from cells under particular conditions onto the microarray. The ‘washing’ process involves the operations of technologically advanced robots in order to ensure that the mRNA is spread evenly over the microarray. If some spots receive more mRNA than others, then misleading data can be generated as to which genes were expressed more than others. Thus, the computer/robotic manipulation required for this step is:

- 2) Washing mRNA over the microarray.

The third step is the image analysis of the microarray. By washing the microarray, the radioactive mRNA has bound to the cDNA spots at different intensities. A robotic ‘eye’ is required in order to measure the intensity of the radioactive glow, which should represent the level of expression of the mRNA in particular cell lines. Once the ‘eye’

measures the intensity, it records that intensity into a computer program, which allows graphs like the one in figure 6.4 to be drawn. Thus, the computer manipulations are the following:

- 3) Robotic 'eye' measures the intensity of the radioactive signal from each cDNA spot.
- 4) The intensity that is recorded is plotted on a graph.

Many statistical tools are used at this point in order to draw out useful information from the data that is presented. Given the large amount of data, which is at least one intensity reading per cDNA spot, specific computer programs are required in order to perform the statistical operations:

- 5) Statistical manipulation of the data.

After the statistical analyses are performed, a researcher may wish to construct interactomes. This is done by clustering the genes that appear to have similar gene expression profiles, and then creating interactomes based on those clusters. Once again, given the complexity of results as seen in figures 6.7, 6.8 and 6.9, programs are required in order to cluster the genes that share these complex gene expression profiles, and then construct interactomes like the one seen in figure 6.10. Therefore, these last two computer manipulations are required:

- 6) Clustering of genes that share expression profiles
- 7) Creating of interactomes based on the clustering.

Lastly, as in the BLAST case, the scientist would analyze the results to determine which conclusions can be drawn, or whether further operations needs to be performed.

- 8) Analysis of the results



The knowledge and operations performed by the human scientist appear very minimal. It seems that the scientists' only task is to decide, once they have a particular problem, which specimens they need to compare, and the final analysis of the results. To give a better sense of which jobs are performed by whom, the steps that are performed using computer databases and programs are listed below, as well as the jobs performed by the human scientist (Table 6.2).

**Table 6.2.** Steps in microarray tests by a human scientist (column 3) and those that are performed using computer databases and programs (columns 1 and 2).

<b>Computer Databases Used</b>	<b>Programs Used</b>	<b>Human operations</b>
Genetic databases of particular organisms. cDNA locations on the microarray.	cDNA spotting on microarray. MRNA washing over cDNA spots. Robotic 'eye' measures the intensity of radioactive signal from each spot. Radioactive intensity plotted onto a graph. Statistical manipulation of data. Clustering of genes that share expression profiles. Creation of interactomes based on the clustering.	Choice of specimens to compare. Monitoring of each computational step. Analysis of the results.

This analysis shows that results from microarrays are not possible without the help of computer databases, programs and robotic mechanisms. What is even more significant is that the results are not always easily understandable by human researchers. For example, the mechanism generated in the antigen example is relatively complex. Other mechanisms have been generated that are even more complex, and they are stored

in computer databases themselves in order to be manipulated in further experiments. Examples of this will be shown later in the section on bioinformatics and interactomes.

**6.8 Analogical Reasoning in Microarray Studies.** Like the examples seen with BLAST and phylogenetic trees, analogical reasoning occurs in some instances of microarray studies but not all. In all microarray studies, including those that study gene expression, classify cancers or discover new diseases, comparisons are made, but not the kind found in analogical reasoning. Analogical reasoning is found in one kind of study that I have presented above, namely the annotation of a new virus by comparing it to known viruses. We will look at both types of studies.

In the case of finding differences in gene expression in cells, often two cell lines are compared. For example, the expression of cells in the human brain can be compared to the cells in the human liver in order to discover what genes are characteristic to brain cells when compared to liver cells. The experiment, in this case, would begin with a microarray with genes from the human genome plated, and then the mRNA products from liver and brain cells, which are radioactively labeled with differently coloured dyes, are washed over the microarray surface. The amount of bound mRNA to the cDNA gives measures as to the expression of particular genes in each cell line, thus giving an indication of how human brain cells are different from human liver cells.

This case, as well as others like it, is unlike the other cases of analogical reasoning we have discussed for the following reason: the needed information about gene expression from the two lines that are being compared is information on how the two

lines are *unlike* each other rather than how they are similar. In the analogical exemplars that we have seen so far, the comparisons are being made in order to find the similarities between the compared phenomena. For example, sound waves are hypothesized to be similar to water waves and the analogy being made between the two of them is for the purposes of finding the similarities that they share. In the case of human brain and liver cells the comparison is being made to see how they are different.

In this case there is still a comparison being made. However, in order to show exactly why it is not an analogy, I will attempt to make this case fit into the multi-constraint schema. Therefore, despite the differences presented above, it may be fruitful to compare the two cases of this example (human brain and liver cells) in Shelley’s (2003) multi-constraint theory schema to see what aspects of analogies are still present in this case, and which are missing (Table 6.3).

**Table 6.3.** Multi-constraint table comparing gene expression on a microarray

<b>Human brain cells</b>	<b>Human liver cells</b>
Human genome	Human genome
Labeled mRNA <sub>b</sub>	Labeled mRNA <sub>l</sub>
Microarray spots	Microarray spots
Spot reader	Spot reader
Express <sub>b</sub> (genome, mRNA <sub>b</sub> )	Express <sub>l</sub> (genome, liver mRNA <sub>l</sub> )
Wash <sub>b</sub> (mRNA <sub>b</sub> , microarray spots)	Wash <sub>l</sub> (mRNA <sub>l</sub> , microarray spots)
Readings <sub>b</sub> (reader, spots)	Readings <sub>l</sub> (reader, spots)
Enable <sub>b0</sub> (express <sub>b</sub> , wash <sub>b</sub> )	Enable <sub>l0</sub> (express <sub>l</sub> , wash <sub>l</sub> )
Enable <sub>b1</sub> (wash <sub>b</sub> , readings <sub>b</sub> )	Enable <sub>l1</sub> (wash <sub>l</sub> , readings <sub>l</sub> )

The table above shows the only similarities between the two cells under comparison. The important part of the comparison, however, is not included, which is comparing the

expression readings from the brain cells to the expression readings from the liver cells. With this important step lacking, it is safe to conclude that the comparison is not analogical.

One class of microarray comparisons is analogical, however. The SARS case does not measure difference in gene expression among cell lines, but instead tries to find genomes that are the most similar. As a reminder, a number of viral genomes were plated on a microarray, and bits of the SARS genome were washed over the microarray in order to find which previously known virus it most closely resembled. Thus, since similarity among genomes is sought, it is likely that the SARS case is an instance of analogical reasoning. The case is put into Table 6.4 below to see if it fits Shelley’s multi-constraint schema.

**Table 6.4.** Multi-constraint table comparing an unknown viral genome to known viral genomes.

<b>Unknown Viral Genome</b>	<b>Known Viral Genomes</b>
Viral genome <sub>u</sub>	Viral genome <sub>k</sub>
Labeled oligonucleotides <sub>u</sub>	Labeled oligonucleotides <sub>k</sub>
Microarray spots	Microarray spots
Spot reader	Spot reader
Portion <sub>u</sub> (genome <sub>u</sub> , oligonucleotides <sub>u</sub> )	Portion <sub>k</sub> (genome <sub>k</sub> , oligonucleotides <sub>k</sub> )
Wash <sub>u</sub> (oligonucleotides <sub>u</sub> , spots)	Wash <sub>k</sub> (oligonucleotides <sub>k</sub> , spots)
Readings <sub>u</sub> (reader, spots)	Readings <sub>k</sub> (reader, spots)
Enable <sub>u0</sub> (portion <sub>u</sub> , wash <sub>u</sub> )	Enable <sub>k0</sub> (portion <sub>k</sub> , wash <sub>k</sub> )
Enable <sub>u1</sub> (wash <sub>u</sub> , readings <sub>u</sub> )	Enable <sub>k1</sub> (wash <sub>k</sub> , readings <sub>k</sub> )

All of the important steps of the viral comparison are presented in the table above. The recorded readings from the unknown viral genome are the same as the readings from

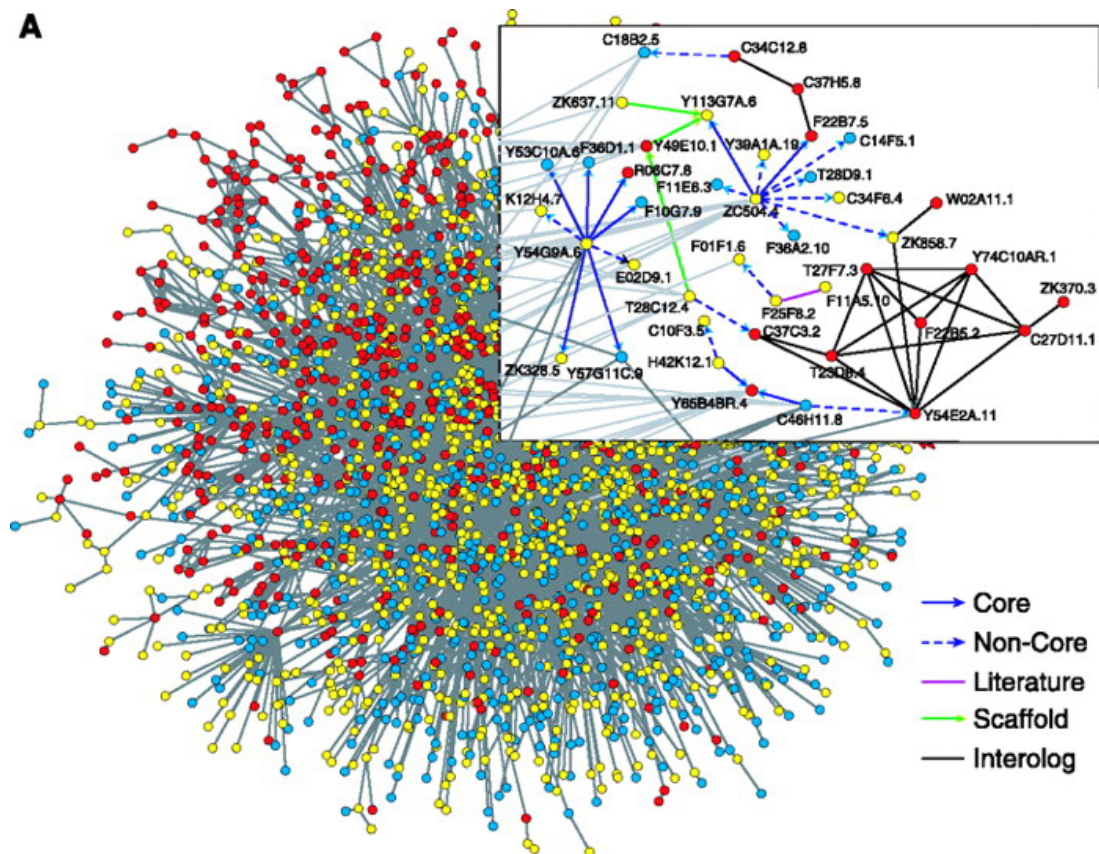
the known viral genome, thus making the two genomes analogous. Thus, microarrays do employ analogical reasoning in some cases.

It is interesting that, in addition to the analogical reasoning that is performed using microarrays, there are analogical studies performed on interactomes, which are the end results of microarray studies. Huang et. al (2004) created a database called POINT (Prediction of Interactome Database, accessible at <http://point.nchc.org.tw:3333/>) that allows scientists to enter protein sequences and the program returns interactomes that similar proteins take part in. This is therefore another example into how analogical reasoning penetrates almost every aspect of bioinformatics study.

It is also apparent in the analogical reasoning performed using microarrays that this reasoning required extended cognition through computer use. The inferences that are made, such as the possible effect of a drug on a virus based on the effect of that drug on a similar virus, are possible only through the use of computers. The computer applications and programs detect and statistically determine the similarities between the entities under study, and if there is a statistical similarity, then a researcher can decide to make the analogical inferences. As seen with BLAST and the creation of phylogenetic trees, bioinformatics research has a unique methodology whereby extended cognition and analogical reasoning are combined.

**6.9 Microarrays and Mechanisms.** One of the most interesting observations about microarray studies is the final result produced. As we saw earlier, the final result of many microarray studies is not only measuring differences in gene expression, but also to create hypothetical interactomes based on the cluster profiles generated from the measurements

of gene expression. Microarrays are becoming very popular due to their ability to quickly and reliably generate these interactomes. This fact is important since interactomes are classic examples of representations of mechanisms. To review, mechanisms have entities and activities that produce regular changes from start up to termination or from within a feedback loop. We saw one example of an interactome with the study on T-cells, and the mechanism is shown in figure 6.10. Many other examples abound, such as the one seen below in figure 6.14, from Li et. al. (2004). This latter figure also shows that the complexity of these interactomes requires extended cognition with computers in order to store and process them.



**Figure 6.14.** The *C. elegans* interactomes. This interactomes has about 4000 interactions between genes, proteins, and other biological molecules (from Li et. al., 2004).

This interactome shows the processes that are working within *C. elegans* cells. Knowledge of these mechanisms can serve many different functions. First, according to the definition presented in Chapter 2, these representations of mechanisms show the entities and activities that occur within the cell. If a scientist were to ask, on a cellular level, what differentiated a stem cell from a fully formed pancreas cell, then the mechanisms at work within the cells would be compared. Second, these mechanisms can help medically in determining what is occurring at a molecular level in medical situations such as the infection.

It may strike one after seeing an interactome such as the one in figure 6.14 that any analysis of these would be next to impossible. This is a criticism of high-throughput sciences in general, which is that these sciences produce volumes of data that are not analyzable. Although I am sure that this is true in many cases, I doubt that it will be a crippling problem. Just as our ability to generate and gather data increases, our ability to analyze such data will also increase. Many in the 16<sup>th</sup> century had been intimidated by the amount and accuracy of Tycho Brahe's astronomical readings, similar to how many are now intimidated by the volume and accuracy of data being produced in bioinformatics research. However, Brahe's readings were used to accurately describe the motions of the heavens, and bioinformatics research will soon accurately describe the workings of life.

A unique aspect of the use of these representations in bioinformatics is that they are readily created by various computer programs. Programs have been developed that can easily create these representations of mechanisms after inputting the analyzed data. This combination of computer use and representations of mechanisms may be a unique

aspect of bioinformatics research. There are clear advantages to this combination, as we will see in the next section.

**6.10 Epistemic Appraisal of Microarrays.** The biological and medical advantages of using microarrays are apparent when one uses Goldman (1992) and Thagard's (1997) appraisal standards for epistemic practices:

*Reliability:* Experiments using microarrays are often beset with numerous problems due to the fact that a large portion of these experiments includes spotting on microarrays, which, by definition, is a 'wet-lab' procedure. These problems range from using impure cDNA, unequal spotting of the cDNA, incomplete absorption of the dyes by the cDNA, and irregularities in fluorescent detection. However, most research using microarrays does take into account these possible problems, and researchers are often confident that they can be overcome using particular statistical techniques. However, the 'dry-lab' steps in microarray research are as reliable as the other bioinformatics tools we have seen so far. Microarrays employ algorithms that perform statistical operations and create complex interactomes, and these programs are reliable scientific tools for generating scientific results.

*Power:* I have demonstrated the popularity of microarrays in a number of different studies. Not only are they used to increase biological knowledge into the inner molecular-biological workings of a cell, but they also has been found to have important medical uses, such as classifying cancers and quickly classifying and potentially finding cures for newly discovered diseases. Due to the popularity of microarrays in the biological and



medical discipline, it is safe to conclude that biologists and medical researchers trust these programs to generate reliable results to many of the questions that interest them.

*Fecundity:* As presented in figures 6.4, 6.5, 6.7, 6.8, 6.13, and 6.14, there is a plethora of results returned in the use of microarrays in bioinformatics for many researchers. Figures 6.4 and 6.5 are the graphs that compare the expression levels of the cell lines compared in the microarray. Figure 6.7 is an array of comparative expressions levels, while figure 6.8 is the cluster analysis of these levels. Figure 6.13 is a compilation of all these figures. Lastly, figure 6.14 is an interactome that is created using the results from numerous interactome studies. As one can see, there is an abundance of results generated, which can only be analyzed using specifically designed algorithms. Just as with BLAST, there is often too much information produced, and the results from just one microarray can generate a large amount of results.

*Speed:* Setting up and analyzing the results from a microarray study can be a very daunting task, at least when compared to studies using BLAST or software that create phylogenetic trees. A researcher needs to order the microarrays, have the proper cDNA spotted, 'wash' the microarray with the dyes, and finally analyze the fluorescent spots. These steps can take many days to weeks to perform. However, microarrays are still much faster than preceding methods in comparing expression levels among cells lines, in creating complex interactomes, and in identifying cancers and viruses. If we take virus identification as an example, previous methods would need to sequence the genome, input that sequence into a genome, and then compare that sequence to other known genomes. This process, as mentioned before, takes weeks as opposed to the microarray process, which only takes a few days.

*Efficiency:* One of the largest hindrances in microarray studies is that the cost of each microarray can be anywhere between \$350 and \$700. Thus, an experiment involving the number of microarrays seen in figure 6.3 can cost up to \$35,000. This figure does not include the instruments that are needed to spot the microarrays with the specific sequences, analyze the radioactive intensities, and so on. Thus, analyses using microarrays are expensive. I have not made a comparison of the cost of microarrays to the cost of other methods, so I cannot comment on whether this method is more cost-effective. However, the increased reliability, power and fecundity of the results that are produced using microarrays may easily outweigh the costs.

*Explanatory Efficacy:* Using microarrays, especially in the creation of complex interactomes, greatly helps in increasing the explanatory efficacy and explanatory coherence of molecular biological studies. By creating interactomes that describe molecular interactions, microarrays help researchers in explaining the molecular workings of organisms. Microarrays are used to create interactomes, interactomes are representations of mechanisms, and most explanations in biology involve describing the entities and activities within mechanisms. Also, since microarrays are able to study entire genomes, researchers are potentially able to explain all the molecular interactions that exist within particular organisms, which lead to explanatory coherence.

**6.11 Summary.** Microarrays, and the programs that analyze their results, are powerful bioinformatics tools that are able to efficiently compare the expression of genes among different cell lines, as well as perform specific operations, such as accurately classifying different types of cancers and identifying novel viruses. In this chapter, I described the

scientific methods that are relevant in research that utilizes microarrays. First, microarray analysis can be seen as an example of extended cognition through computer use because the size of the microarrays that are made, the different expression levels that are generated by the spots on the microarray, the analysis of those spots, the creation of interactomes, and many other specific operations. Second, the use of microarrays can be seen as an example of analogical reasoning, as in the discovery of new viruses by comparing their sequences to the sequences of viral species that are already known. Third, the interactomes that are created from microarray analyses are examples of representations of mechanisms. Fourth, research with microarrays combines the methods of computer use and analogical reasoning with the creation of representations of mechanisms. Lastly, the use of microarrays meets all the standards stated by Goldman (1992) and Thagard (1997) for the appraisal of epistemic practices.

## Chapter 7

# Conclusion

**7.1 Introduction.** This thesis has discussed various scientific methods that are relevant to bioinformatics research. These methods include extended cognition, analogical reasoning, and the creation of representations of mechanisms for scientific explanation, as well as various combinations of these methods. By analyzing specific tools in bioinformatics research, such as the use of BLAST, phylogenetic tree creation, and the use of microarrays, I have tried to show this relevance. Lastly, I have performed an epistemic appraisal of these bioinformatics tools using Goldman (1992) and Thagard's (1997) standards. I will summarize each of these points below.

**7.2 Extended Cognition and Bioinformatics.** Chapters 4, 5 and 6 demonstrated how extended cognition occurs when scientists use bioinformatics tools like BLAST, microarrays and phylogenetic analysis software. In each of these cases, the main role of scientists themselves is to decide what kind of analysis is to be performed and a final analysis of the results. For example, while using BLAST, the scientist decides which sequences to compare, and while using microarrays, which cell lines to compare. Once these decisions are made, then the requisite information is extracted from various computer databases and specific computer algorithms perform the requisite operations. Thus, any hypotheses generated using these procedures are examples of extended cognition.

One should not conclude that this means that the role of the scientist has been marginalized, however. Analysis of results, which is the final step in all the bioinformatics research presented in this thesis, is a major step in all scientific research. Although software has been developed to analyze data, some human analysis is still necessary. Also, there is certainly a major component required in all of scientific research, which is *creativity*. The development of tools such as BLAST and microarrays, and using these tools in novel ways, such as using microarrays for viral discovery, is a creative human process. There are still many operations performed by humans that are not even close to being mimicked or replaced by computer software.

One fear that arises from any overuse of extended cognition is that researchers will no longer understand the operations being accomplished by their research partners or by the computer software. There are two responses to this fear. The first is that this gap in understanding is not necessarily new in science. For example, many biologists use electron microscopes without having an understanding of how the wavelengths of electrons contribute to observing smaller particles. Social scientists use statistical software or employ statisticians to analyze their data without understanding the details of normal bell curves, ANOVA, regressions, or any of the other statistical concepts and tools being employed. The second response is that if one wanted to gain the understanding they are lacking, they can simply ask. If the biologists wished to know the workings of an electron microscope, they can ask an engineer or physicist familiar with the technology. If a biologist wished to learn how BLAST worked, they can ask a computer scientist who specializes in bioinformatics.

**7.3 Analogical Reasoning and Bioinformatics.** Bioinformatics research is very dependent upon analogical reasoning, since bioinformatics specializes in comparing biological data to one another, and the properties or actions of one set of data are shown to be similar to the properties or actions of the other set. In BLAST, query sequences are compared to a database of sequences. BLAST returns target sequences that are most similar to the query sequence and the former sequences are hypothesized to have similar properties to the latter sequences. With microarrays, viral sequences are compared to one another and the ones that are most closely matched on the microarrays are hypothesized to share the most properties. In the creation of phylogenetic trees, larger sequences, possibly even genomes are compared. The more closely related these sequences are to each other, the more likely the species are closely related to one another. Using Gentner (1983) and Shelley's (2002) structure-mapping method for mapping analogies, I have shown how each of these comparisons is an example of an analogy.

What is interesting about bioinformatics research, however, is that this analogical reasoning is performed with the help of computer use. This use is somewhat unique in bioinformatics, but there are other modern scientific fields that are also employing this combination. For example, climate scientists can create simulations of weather systems with the intention that these simulations are analogical to actual weather systems. The increased use of computers in scientific fields may allow for these types of analogical reasoning to become more prevalent, and thus further supporting the scientific merit of this type of reasoning.

**7.4 Representations of Mechanisms and Bioinformatics.** Although the use of BLAST and the creation of phylogenetic trees do not directly produce examples of representations of mechanisms, the use of microarrays does produce such examples. By comparing the expression levels of cell lines, researchers can get an idea of which molecules interact with one another. This interaction allows researchers, or more precisely, specifically designed software, to create complex representations of mechanisms. These representations, which are often called ‘interactomes’, help biological researchers understand the molecular underpinnings of organisms, and help medical researchers create better treatments for particular diseases.

Since these representations are created using specifically designed programs, bioinformatics research may be unique in combining the methodologies of computer use and the creation of representations of mechanisms. However, just as with the combination of computer use and analogical reasoning, this new combination may also be used in other scientific fields. Fields such as ecology, climatology, and even physics may be able to use computers to generate complex representations of mechanisms after collecting and compiling various data.

**7.5 Epistemic Appraisal of Bioinformatics.** Using Goldman (1992) and Thagard’s (1997) standards of epistemic appraisal, I have shown how each of the cases presented in the previous three chapters fit these standards. With each of these cases fitting the standards, one can potentially give an epistemic appraisal of the whole field of bioinformatics. However, because the tools presented in chapters 4, 5 and 6 are seen as

the major tools employed in bioinformatics research, thus one can safely conclude that bioinformatics as a field meets these epistemic standards.

**7.6 Future Prospects.** As a field, bioinformatics is quickly growing, with many new developments that continue to astound the scientific community. If this thesis had been written ten years ago, then the topic of representations of mechanisms would probably not have been presented to be relevant to bioinformatics, since microarray research was still in its infancy.

Bioinformatics continues to grow with more databases, more precise algorithms, and algorithms that are designed to find different types of information. For example, Magdaleno et. al. (2006) created the Development of the Brain Gene Expression Map which gives precise information on the gene expression of the human brain. Mayer et. al. (2005) developed an algorithm that is able to find differences between sequences in their conserved domains (recall DNA domains from section 3.6). Oldham et. al. (2006) have used microarrays to find the differences in gene expression between humans and chimpanzees, helping to further characterizes the differences between our two species. More radical developments include Mastrobattista et. al's (2005) Water-oil-water (WOW) test, which is able to identify proteins within a matter of days rather than months. Haque et. al.'s (in review) 'biochips' are able to generate 60 times more data than current technological methods by reading ions across a cell's membranes, where the cells are placed in tiny pockets inside specially designed microchips. Meanwhile, the number of sequenced genes, proteins, and genomes continues to grow at an exponential



rate. I hope that this thesis provides a basic philosophical understanding of the field of bioinformatics.

## Reference List

1. Adams MD et. al. (list of authors takes up a page) (2000) “The genome sequence of *Drosophila melanogaster*” *Science*. 287:2185-95.
2. Alberts B. (2002) *Molecular Biology of the Cell*, Garland Publishing.
3. Altschul S.F., Gish W., Miller W., Myers E.W. and Lipman D.J. (1990) “Basic local alignment search tool” *Journal of Molecular Biology*. 215:403-410.
4. Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) “Gapped BLAST and PSI-BLAST: a new generation of protein database search programs” *Nucleic Acids Research*. 25:3389-3402.
5. Alvarez L.W., Alvarez W., Asaro F. and Michel H.V. (1980) “Extraterrestrial cause for the Cretaceous-Tertiary extinction: Experimental results and theoretical interpretation” *Science*. 208:1095-1108.
6. Aristotle (1991) *The History of Animals*, translated by DM Balme, Cambridge, MA, Harvard University Press.
7. Aristotle (1997) *Topics*, translated by R Smith, Oxford, NY, Oxford University Press.
8. Ashburner M., Drysdale R. (1994) “FlyBase - the *Drosophila* genetic database” *Development*. 120:2077-9.
9. Ayer A.J. (1952) *Language, Truth and Logic*, Dover, New York, NY.
10. Baldi P. and Hatfield G.W. (2002) *DNA Microarrays and Gene Expression: From Experiments to Data Analysis and Modeling*, Cambridge University Press, Cambridge, UK.

11. Bechtel W. (1994) "Levels of Description and Explanation in Cognitive Science" *Minds and Machines: Journal for Artificial Intelligence, Philosophy, and Cognitive Science*. 4:1-25.
12. Bechtel W. and Abrahamsen A. (2005). "Explanation: A Mechanistic Alternative." *Studies in History and Philosophy of the Biological and Biomedical Sciences*. 36: 421-441.
13. Benson D.A., Karsch-Mizrachi I., Lipman D.J., Ostell J., Wheeler D.L. (2005) "GenBank" *Nucleic Acids Research*, 33(Database issue):D34-38.
14. Berman H.M., Westbrook J., Feng Z., Gilliland G., Bhat T.N., Weissig H., Shindyalov I.N., Bourne P.E. (2000) "The Protein Data Bank" *Nucleic Acids Research*. 28: 235-242.
15. Blanchet C., Combet C., Geourjon C., Deleage G. (2000) "MPSA: integrated system for multiple protein sequence analysis with client/server capabilities" *Bioinformatics*. 16: 286-7.
16. Bonner J.T. (1963) "Analogies in Biology" *Synthese: An International Journal for Epistemology, Methodology and Philosophy of Science*. 15: 275-279.
17. Bowcock A.M., Ruiz-Linares A., Tomfohrde J., E. Minch, J.R. Kidd, and L.L. Cavalli-Sforza (1994) "High resolution of human evolutionary trees with polymorphic microsatellites" *Nature*. 368: 455-457.
18. Brewster J.L., K. Beth Beason, Todd T. Eckdahl, and Irene M. Evans (2004) "The Microarray Revolution" *Biochemistry and Molecular Biology Education*. 32:217-227.
19. Brown W.R. (1989) "Two Traditions of Analogy" *Informal Logic*. 11: 161-172.

20. *C. elegans* Sequencing Consortium (1988) "Genome sequence of the nematode *C. elegans*: a platform for investigating biology." *Science*. 282:2012-8.
21. Cann R.L., M. Stoneking and A. C. Wilson (1987) "Mitochondrial DNA and Human Evolution" *Nature*. 325: 31-36.
22. Clark A. and D.J. Chalmers (1998) "The Extended Mind" *Analysis*. 58:10-23
23. Crouau-Roy B., Service, S., Slatkin M. and Freimer N. (1996) "A fine-scale comparison of the human and chimpanzee genomes: linkage, linkage disequilibrium and sequence analysis" *Human Molecular Genetics*. 5:1131-7.
24. Darden, L. (1998) "Recent Work in Computational Scientific Discovery" in *Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society*. Michael Shafto and Pat Langley (Eds.). Mahwah, New Jersey: Lawrence Erlbaum, pp. 161-166.
25. Darwin C. (1859) *On the Origin of Species Means of Natural Selection: Or the Preservation of Favoured Races in the Struggle for Life*. London, reprinted with an introduction by Ernst Mayr. Harvard University Press, Cambridge, MA.
26. Dayhoff M.O. (1965) *Atlas of Protein Sequences*. National Biomedical Research Foundation, Silver Spring, Md.
27. Dayhoff M.O. (1978) "Survey of new data and computer methods of analysis" In M. O. Dayhoff, ed., *Atlas of Protein Sequence and Structure*, 5:29, National Biomedical Research Foundation, Silver Springs, Md.
28. Dickerson R.C. (1971) "The structure of cytochrome C and the rates of molecular evolution" *Journal of Molecular Evolution*. 1:26-45

29. Dixon R.A., Kobilka B.K., Strader D.J., Benovic J.L., Dohlman H.G., Frielle T., Bolanowski M.A., Bennett C.D., Rands E., Diehl R.E., et al. (1986) "Cloning of the gene and cDNA for mammalian beta-adrenergic receptor and homology with rhodopsin" *Nature*. 321:75-79
30. Doolittle, R.F. (1981) "Similar amino acid sequences: chance or common ancestry?" *Science*, 214:149-59.
31. Downward J., Yarden Y., Mayes E., Scrace G., Totty N., Stockwell P., Ullrich A., Schlessinger J. and Waterfield M.D. (1984) "Close similarity of epidermal growth factor receptor and *v-erb-B* oncogene protein sequences" *Nature*. 307:521-527.
32. Durbin R., S. R. Eddy, A. Krogh, G. Mitchison (1999) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge, UK.
33. Eck R.V., Dayhoff M.O. (1966) *Atlas of Protein Sequence and Structure*. National Biomedical Research Foundation, Silver Spring, MD.
34. Edgar R., Domrachev M., Lash A.E. (2002) "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository" *Nucleic Acids Research*, 30, 1:207-210.
35. Efron B., Tibshirani R., Storey J.D., Tusher V. (2001) "Empirical Bayes Analysis of a Microarray Experiment" *Journal of the American Statistical Association*. 96: 1151-1160.
36. Falcon A. (1996) "Aristotle's Rules of Division in the *Topics*: The Relationship between Genus and Differentia in a Division" *Ancient Philosophy*. 16: 377-387.

37. Fitch W.M. (1977) "On the problem of discovering the most parsimonious tree" *American Naturalist*. 111: 223-257.
38. Fleischmann R.D., Adams M.D., White O., Clayton R.A., Kirkness E.F., Kerlavage A.R., Bult C.J., Tomb J.F., Dougherty B.A., Merrick J.M., et al. (1995) "Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd.", *Science*. 269:496-512.
39. Feng D. and R. F. Doolittle (1987) "Progressive sequence alignment as a prerequisite to correct phylogenetic trees." *J. Mol. Evol.* 60:351-360.
40. Galibert F., Alexandraki D., Baur A., Boles E., Chalwatzis N., Chuat J.C., Coster F., Cziepluch C., De Haan M., Domdey H., Durand P., Entian K.D., Gatus M., Goffeau A., Grivell L.A., Hennemann A., Herbert C.J., Heumann K., Hilger F., Hollenberg C.P., Huang M.E., Jacq C., Jauniaux J.C., Katsoulou C., Karpfinger-Hartl L., et al. (1996) "Complete nucleotide sequence of *Saccharomyces cerevisiae* chromosome X." *European Molecular Biology Organization Journal*. 15:2031-49.
41. Giere R. (1988) *Explaining Science: A Cognitive Approach*, University of Chicago Press, Chicago.
42. Giere R. (1999) "Using Models to Represent Reality" In *Model-Based Reasoning in Scientific Discovery*, Ed. L. Magnani, N. J. Nersessian, and P. Thagard, Kluwer/Plenum, New York.
43. Giere R. (2002a) "Scientific Cognition as Distributed Cognition" In *Cognitive Bases of Science*, eds. Peter Carruthers, Stephen Stich and Michael Siegal, Cambridge: Cambridge University Press.

44. Giere R. (2002b) Models as Parts of Distributed Cognitive Systems, In *Model Based Reasoning: Science, Technology, Values*, 227-41, eds. Lorenzo Magnani and Nancy Nersessian, Kluwer.
45. Giere R., Moffatt B. (2003) "Distributed Cognition: Where the Cognitive and the Social Merge" *Social Studies of Science*. 33: 301-310
46. Giere R. (2004) "The Problem of Agency in Scientific Distributed Cognitive Systems" *Journal of Cognition and Culture*. 4: 759-74.
47. Gentner D. (1983). "Structure-mapping: A theoretical framework for analogy" *Cognitive Science*. 7: 155-170.
48. Goldman A.I. (1992) *Liaisons: Philosophy meets the cognitive and social sciences*, MIT Press, Cambridge, MA.
49. Golub T.R., Slonim D.K., Tamayo P., Huard C., Gaasenbeek M., Mesirov J.P., Coller H., Loh M.L., Downing J.R., Caligiuri M.A., Bloomfield C.D., Lander E.S. (1999) "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring Science" *Science*. 286: 531-537.
50. Goodman M., Moore G.W., Barnabas J., Matsuda G. (1974) "The phylogeny of human globin genes investigated by the maximum parsimony method" *Journal of Molecular Evolution*. 3:1-48.
51. Griffiths A.J.F. (2002) *Modern Genetic Analysis*, W.H. Freeman & Company, New York, NY.
52. Guermeur Y., Geourjon C., Gallinari P., Deleage G. (1999) "Improved Performance in Protein Secondary Structure Prediction by Inhomogeneous Score Combination" *Bioinformatics*. 15: 413-421.

53. Hager T. (1995) *Force of Nature: The Life of Linus Pauling*, Simon & Schuster.
54. Hall B.G. (2001) *Phylogenetic Trees Made Easy. A How-To for Molecular Biologists*. Sinauer Associates, Sunderland, MA.
55. Hamming R.W. (1950) "Error-detecting and error-correcting codes" *Bell System Technical Journal*. 29: 147-60.
56. Haque A.U., Rokkam M., De Carlo A.R., Wereley S.T., Wells H.W., McLamb W.T., Roux S.J., Irazoqui P.P., Porterfield D.M. (In review) "A MEMS Fabricated Cell Electrophysiology Laboratory Biochip for *In-silico* Calcium Measurements." *Sensors and Actuators*.
57. Harding (2005) "BLAST" *The Scientist*. 19: 21.
58. Hesse, M. (1952) "Operational Definition and Analogy in Physical Theories" *British Journal for the Philosophy of Science*. 2: 281-294.
59. Hesse, M. (1966) *Models and Analogies in Science*. University of Notre Dame Press, Notre Dame.
60. Holyoak & Thagard (1995) *Mental Leaps: Analogy in Creative Thought*, MIT Press, Cambridge, MA.
61. Huang T.W., Tien A.C., Huang W.S., Lee Y.C., Peng C.L., Tseng H.H., Kao C.Y., Huang C.Y. (2004) "POINT: a database for the prediction of protein-protein interactions based on the orthologous interactome" *Bioinformatics*. 20:3273-6.
62. Huelsenbeck J. P. and Ronquist F. (2001) "MRBAYES. Bayesian inference of phylogeny" *Bioinformatics*, 17:754-755.
63. Humphreys P. (2004) *Extending Ourselves*, Oxford University Press, Oxford, UK.
64. Hutchins E. (1995) *Cognition in the Wild*, MIT Press, Cambridge, MA.



65. International Human Genome Consortium (2001) "Initial sequencing and analysis of the human genome," *Nature*, 409: 860-922.
66. Jorde L.B, Bamshad M., Rogers A.R. (1998) "Using mitochondrial and nuclear DNA markers to reconstruct human evolution" *Bioessays*. 20:126-36.
67. Joshi-Tope G., Gillespie M., Vastrik I., D'Eustachio P., Schmidt E., de Bono B., Jassal B., Gopinath G.R., Wu G.R., Matthews L., Lewis S., Birney E., Stein L. (2005) "Reactome: a knowledgebase of biological pathways." *Nucleic Acids Research*. 1:33 (Database issue).
68. Karp P., Ouzouni C. and Paley, S. M. (1996) "HinCyc: A Knowledge Base of the Complete Genome and Metabolic Pathways of *H. influenzae*," in D.J. States, P. Agarwal, T. Gaasterland, L. Hunter & R. Smith (Eds.), *ISMB-96: Proceedings of the Fourth International Conference on Intelligent Systems for Molecular Biology*, pp. 116-124. AAAI Press, Menlo Park, CA.
69. Krebs H.A., Johnson W.A. (1937) "Metabolism of ketonic acids in animal tissues" *Biochemistry Journal*, 31:645-60.
70. Lafolette H. and N. Shanks (1996) *Brute Science - Dilemmas of Animal Experimentation*. Routledge, New York, NY.
71. Lashkari D.A., DeRisi J.L., McCusker J.H., Namath A.F., Gentile C.H., Seung Y. Brown P.O., Davis R.W. (1997) "Yeast microarrays for genome wide parallel genetic and gene expression analysis" *Proceedings of the National Academy of Sciences, USA*. 94:13057-13062.

72. Laskowski R.A., Hutchinson E.G., Michie A.D., Wallace A.C., Jones M.L. and Thornton J.M. (1997) "PDBsum: a Web-based database of summaries and analyses of all PDB structures" *Trends in Biochemistry Science*, 22:488-90.
73. Latour B. (1987) *Science in action: How to follow scientists and engineers through society*. Harvard University Press, Cambridge, MA.
74. Lee D.S. (1969) "Analogy in Scientific Theory Construction" *Southern Journal of Philosophy*. 7: 107-125.
75. Leung Y.F. and D. Cavalieri (2003) "Fundamentals of cDNA microarray data analysis" *Trends in Genetics*. 19: 649-659.
76. Lewin B. (1997) *Genes VI*. Oxford University Press, New York, NY.
77. Li S., C.M. Armstrong, N. Bertin, H. Ge, S. Milstein, M. Boxem, P. Vidalain, J. J. Han, A. Chesneau, T. Hao, D.S. Goldberg, N. Li, M. Martinez, J. Rual, P. Lamesch, L. Xu, M. Tewari, S.L. Wong, L.V. Zhang, G.F. Berriz, L. Jacotot, P. Vaglio, J. Reboul, T. Hirozane-Kishikawa, Q. Li, H.W. Gabel, A. Elewa, B. Baumgartner, D.J. Rose, H. Yu, S. Bosak, R. Sequerra, A. Fraser, S.E. Mango, W.M. Saxton, S. Strome, S. van den Heuvel, F. Piano, J. Vandenhoute, C. Sardet, M. Gerstein, L. Doucette-Stamm, K.C. Gunsalus, J.W. Harper, M.E. Cusick, F.P. Roth, Hill D.E., M. Vidal (2004) "A Map of the Interactome Network of the Metazoan *C. elegans*", *Science* 303:540-543.
78. Lipman D., Maizel J. (1982) "Comparative analysis of nucleic acid sequences by their general constraints" *Nucleic Acids Research*. 10: 2723-39.
79. Lipman D.J., Pearson W.R. (1985) "Rapid and sensitive protein similarity searches." *Science*. 227:1435-41.

80. Lu D., Chen S., Zhang S., Zhang M., Zhang W., Bo X., Wang S. (2005) "Screening of specific antigens for SARS clinical diagnosis using a protein microarray" *Analyst*. 130: 474-82.
81. Lwoff A. and Ullmann, A. (1979) *Origins of Molecular Biology. A Tribute to Jacques Monod*. Academic Press, New York, NY.
82. Machamer P., Darden L., Craver C.F. (2000) "Thinking about mechanisms" *Philosophy of Science*. 67: 1-25.
83. Magdaleno S., P. Jensen, C.L. Brumwell, A. Seal, K. Lehman, A. Asbury, T. Cheung, T. Cornelius, D.M. Batten, C. Eden, S.M. Norland, D.S. Rice, N. Dosooye, S. Shakya, P. Mehta, T. Curran (2006) "BGEM: An In Situ Hybridization Database of Gene Expression in the Embryonic and Adult Mouse Nervous System" *Public Library of Science Biology*. 4: 317-328.
84. Mastrobattista E., V. Taly, E. Chanudet, P. Treacy, B.T. Kelly, and A.D. Griffiths (2005) "High-throughput screening of enzyme libraries: *In vitro* evolution of a  $\beta$ -Galactosidase by fluorescence-activated sorting of double emulsions" *Chemistry & Biology*. 12: 1291-1300.
85. Mayer K.M., S.R. McCorkle and J. Shanklin (2005) "Linking enzyme sequence to function using conserved property difference locator to identify and annotate positions likely to control specific functionality" *BMC Bioinformatics*. 6:284.
86. McConkey E. (2002) "Altered gene expression could explain the genetic difference between human and chimp," *The Scientist*, 16.
87. Mill J.S. (1873) *System of Logic*. University Press of the Pacific, San Diego, CA.

88. Needleman, S.B., Wunsch, C.D. (1970) "A general method applicable to the search for similarities in the amino acid sequence of two proteins." *Journal of Molecular Biology*. 48: 443-53.
89. Nei M. (1995) "Genetic absolute dating based on microsatellites and the origin of modern humans" *Proceedings of the National Academy of Science USA*. 92: 6720–6722.
90. O'Brien K.P., I. Westerlund, and E.L.L. Sonnhammer (2004) "OrthoDisease: A Database of Human Disease Orthologs" *Human Mutation*. 24: 112-9.
91. Oldham M.C., Horvath S., Geschwind D.H. (2006) "Conservation and evolution of gene coexpression networks in human and chimpanzee brains" *Proceedings of the National Academy of Science of the U.S.A.* 103: 17973-17978.
92. Pardee A., F. Jacob, and J. Monod (1959) "The genetic control and cytoplasmic expression of 'inducibility' in the synthesis of beta-galactosidase by *E. coli*" *Journal of Molecular Biology*. 1: 165.
93. Pauling, L. (1939) "The structure of proteins" *Journal of the American Chemical Society*. 61: 1860-1867.
94. Peri S., Navarro J.D., Amanchy R., Kristiansen T.Z., Jonnalagadda C.K., Surendranath V., Niranjana V., Muthusamy B., Gandhi T.K., Gronborg M., Ibarrola N., Deshpande N., et. al. (2003) "Development of human protein reference database as an initial platform for approaching systems biology in humans" *Genome Research*. 13:2363-2371.

95. Perutz M.F., Kendrew J.C. and Watson H.C. (1965) "Structure and function of haemoglobin II. Some relations between polypeptide chain configuration and amino acid sequence" *Journal of Molecular Biology*. 13: 669-78.
96. Pevsner J. (2003) *Bioinformatics and Functional Genomics*. John Wiley & Sons, Hoboken, NJ.
97. Pinker S. (2003) *The Blank Slate: the modern denial of human nature*. Penguin, Toronto, ON.
98. Potti A., Dressman H.K., Bild A., Riedel R.F., Chan G., Sayer R., Cragun J., Cottrill H., Kelley M.J., Petersen R., Harpole D., Marks J., Berchuck A., Ginsburg G.S., Febbo P., Lancaster J., Nevins J.R. (2006) "Genomic signatures to guide the use of chemotherapeutics" *Nature*. 12: 1294-1300.
99. Qiu M., Shi Y., Guo Z., He R., Chen R., Zhou D., Dai E., Wang X., Si B., Song Y., Li J., Yang L., Wang J., Wang H., Pang X., Zhai J., Du Z., Liu Y., et al. (2005) "Antibody response to individual proteins of SARS coronavirus and their naturalization activities" *Microbes and Infection*. 7: 882-9.
100. Quartz S., Sejnowski T.J. (1997) "The neural basis of cognitive development: A constructivist manifesto," *Behavioral & Brain Sciences*, 20: 537-596.
101. Ridley, M (1996) *Evolution*. Blackwell Science, Cambridge, MA.
102. Rubin G.M., M.D. Yandell, J.R. Wortman, G.L. G. Miklos, C.R. Nelson, I.K. Hariharan, M.E. Fortini, P.W. Li, R. Apweiler, W. Fleischmann, J.M. Cherry, S. Henikoff, M.P. Skupski, S.Misra, M. Ashburner, E. Birney, M.S. Boguski, T. Brody, P. Brokstein, S.E. Celniker, et. al. (2000) "Comparative Genomics of the Eukaryotes" *Science*. 287: 2204-15.

103. Rupert R.D. (2004) "Challenges to the hypothesis of extended cognition" *The Journal of Philosophy*. 101: 389-428.
104. Ruse M. (1973) "The value of analogical models in science" *Dialogue: Canadian Philosophical Review*. 12: 246-253.
105. Sanger, F., Nicklen, S., Coulson, A.R. (1977) "DNA sequencing with chain-terminating inhibitors." *Proceedings of the National Academy of Science USA*. 74: 5463-7.
106. Sanger F., Coulson A.R., Hong G.F., Hill D.F. and Petersen G.B. (1982) "Nucleotide sequence of bacteriophage lambda DNA." *Journal of Molecular Biology*. 162: 729-773.
107. Sattin B.D., Pelling A.E., Goh M.C. (2004) "DNA base pair resolution by single molecule force spectroscopy" *Nucleic Acids Research*. 2004 32: 4876-4883
108. Schena M., Shalon D., Davis R.W., Brown P.O. (1995) "Quantitative monitoring of gene expression patterns with a complementary DNA microarray." *Science*. 270: 467-70.
109. Shalon M., Davis R.W., Brown P.O. (1995) "Quantitative monitoring of gene expression patterns with a complementary DNA microarray" *Science*. 270: 467-470.
110. Shelley C. (2002) "Analogy counterparts and the acceptability of the analogical hypothesis" *British Journal of Philosophy of Science*. 53: 477-96.
111. Shelley C. (2003) "Multiple Analogies in Science and Philosophy" *Pragmatics and Cognition*, 12.

112. Shelley C. (2004) "Analogy Counterarguments: A Taxonomy for Critical Thinking" *Argumentation: An International Journal on Reasoning*. 18: 223-238.
113. Simon H. A. (1978) "On the forms of mental representation" In C.W. Savage (Ed.), *Perception and cognition: Issues in the foundations of psychology*. 9: 3-18. *Minnesota Studies in the Philosophy of Science*. University of Minnesota Press, Minneapolis.
114. Smith T.F., Waterman M.S. (1981) "Identification of common molecular subsequences." *Journal of Molecular Biology*. 147: 195-7.
115. Sokal R.R., Michener C.D. (1958) "Statistical Methods for evaluating systemic relationships" *University of Kansas Science Bulletin*. 38:1409-38.
116. Southern E.M. (1975) "Detection of specific sequences among DNA fragments by gel-electrophoresis" *Journal of Molecular Biology*. 98:503.
117. Southern E.M. (2001) "DNA Microarrays. History and overview" *Methods in Molecular Biology*. 170: 1-15.
118. Staden R. (1977) "Sequence data handling by computer" *Nucleic Acids Research*. 4: 4037-51.
119. Stein L., Sternberg P., Durbin R., Thierry-Mieg J., Spieth J. (2001) "WormBase: network access to the genome and biology of *Caenorhabditis elegans*" *Nucleic Acids Research*. 29:82-6.
120. Stein L. (2005) "What's next for bioinformatics?" *The Scientist*. 19:31.
121. Stepan N.L. (1986) "Race & Gender: The Role of Analogy in Science" *Critiques & Contentions*. 77:261-277

122. Stoneking M. (1994) "Mitochondrial DNA and human evolution" *Journal of Bioenergy and Biomembranes*. 26: 251-59.
123. Swofford D.L. (1991) *PAUP: Phylogenetic Analysis Using Parsimony*, Macintosh Version 3.0r, Computer program distributed by the Illinois Natural History Survey, Champaign, Illinois.
124. Thagard P. (1988) "Computational models in the philosophy of science" In A. Fine & P.Machamer (Eds.), *PSA 1986*, 2: 329-335). Philosophy of Science Association, East Lansing, MI.
125. Thagard P. (1993) "Societies of minds: Science as distributed computing" *Studies in History and Philosophy of Science*. 24: 49-67.
126. Thagard P. (1997) "Collaborative Knowledge" *Nous*. 31: 242-261.
127. Thagard P. (2003) "Pathways to biomedical discovery" *Philosophy of Science*, 70: 235-254.
128. Thornton J.W., and DeSalle R. (2000) "Gene family evolution and homology: genomics meets phylogenetics" *Annual Review of Genomics and Human Genetics*, 1:41-73.
129. Van Fraassen B. (1980) *The Scientific Image*. Clarendon Press, New York.
130. Venter J.C. (1995) "*E. coli* sequencing." *Science*. 267:601.
131. Von Eckardt B., Poland J. (2004) "Mechanism and Explanation in Cognitive Neuroscience" *Philosophy of Science*. 71:972-984.
132. Wang D., Urisman A., Liu Y.T., Springer M., Ksiazek T.G., Erdman D.D., Mardis E.R., Hickenbotham M., Magrini V., Eldred J., Latreille J.P., Wilson R.K.,



- Ganem D., DeRisi J.L. (2003) "Viral Discovery and Sequence Recovery Using DNA Microarrays" *Public Library of Science Biology*. 1: 257-60.
133. Weinstein J.N., Myers T.G., O'Connor P.M., Friend S.H., Fornace A.J. Jr., Kohn K.W., Fojo T., Bates S.E., Rubinstein L.V., Anderson N.L., Buolamwini J.K., van Osdol W.W., Monks A.P., Scudiero D.A., Sausville E.A., Zaharevitz D.W., Bunow B., Viswanadhan V.N., Johnson G.S., Wittes R.E., Paull K.D. (1997) "An information-intensive approach to the molecular pharmacology of cancer" *Science*. 275: 343-349.
134. Wheeler D.L., Chappey C., Lash A.E., Leipe D.D., Madden T.L., Schuler G.D., Tatusova T.A., Rapp B.A. (2000) "Database resources of the National Center for Biotechnology Information" *Nucleic Acids Research*, 28:10-4.
135. Wilbur, W.J., Lipman, D.J. (1983) "Rapid similarity searches of nucleic acid and protein data banks." *Proceedings of the National Academy of Science USA*. 80:726-30.
136. Wilson M. (1997) "Analogy in Aristotle's Biology" *Ancient Philosophy*. 17: 335-358.
137. Wilson P.R. (1964) "On the argument by analogy" *Philosophy of Science*. 62: 364-73.
138. Wilson R.A. (1994) "Wide Computationalism" *Mind: A Quarterly Review of Philosophy*. 103: 351-372.
139. Wu C.H., Lai-Su L.Y., Huang H., Arminski L., Castro-Alvear J., Chen Y., Hu Z., Kourtesis1 P., Ledley R.S., Suzek B.E., Vinayaka C.R., Zhang J., Barker W.C. (2003) "Protein Information Resource" *Nucleic Acids Research*, 31: 345-7.

140. Yamashita S., Y. Tanaka, S. Tsutsumi, H. Aburatani, N. Minato, S. Ihara (2005)  
“Analysis of mechanism for human  $\gamma\delta$  T cell recognition of nonpeptide antigens”  
*Biochemical and Biophysical Research Communications*. 334: 349–360.