Linköping Electronic Articles in
Computer and Information Science
Vol. 3(1998): nr 7

Uncertainty in AI and Bioinformatics

Frans Voorbraak

Medical Informatics
Academic Medical Center
University of Amsterdam
Meibergdreef 15
1105 AZ Amsterdam, The Netherlands
f.p.voorbraak@amc.uva.nl

Linköping University Electronic Press
Linköping, Sweden

http://www.ep.liu.se/ea/cis/1998/007/

Published on October 15, 1998 by
Linköping University Electronic Press
581 83 Linköping, Sweden

Linköping Electronic Articles in
Computer and Information Science
ISSN 1401-9841
Series editor: Erik Sandewall

Typeset by the author using LaTeX
Formatted using étendu *style*

Recommended citation:
<Author>. <Title>. Linköping Electronic Articles in Computer and Information Science, Vol. 3(1998): nr 7. http://www.ep.liu.se/ea/cis/1998/007/. *October 15, 1998.*

This URL will also contain a link to the author's home page.

Abstract

In recent years, new efficient experimental techniques, especially in the area of DNA sequencing, have led to a tremendous growth in available biological data. Many large sequence databases are already publicly available on the internet, and information is added at a spectacular rate. The extensive human genome project is only one of the many sources of this information.

It is widely recognized that the mere gathering of data is not sufficient and that its biological interpretation is of the utmost importance. Unfortunately, the development of methods for interpreting the data is not keeping up with the tempo with which the data is accumulated.

It is clear that many types of questions can only be asked by a computational analysis, and computer science has become an integral part of the research involving biological sequences (of DNA, RNA, or proteins). The research area combining biology and computer science is known as bioinformatics. Conventional computer methods and algorithms have been applied quite successfully in this area, but the often enormous amounts of data to be analyzed and the complexity of biological systems leave many interesting problems beyond the reach of conventional approaches.

The challenging computational problems of bioinformatics provide interesting opportunities for applying methods from the field of artificial intelligence. In this paper, the emphasis is on discussing how methods from the field of uncertainty in AI can be relevant for some challenging problems of bioinformatics. Some necessary background information on molecular genetics in general and the human genome project in particular is provided at the beginning of the paper.

1 Introduction

In recent years, new efficient experimental techniques have led to a tremendous growth in available biological data. The most obvious sources are the large genome projects and numerous other research efforts contributing to DNA sequencing, which can roughly be described as the process of determining the linear order of the (four) different types of nucleotides within (part of) a DNA-molecule. Many large sequence databases are already publicly available on the internet, and information is added at a spectacular rate. But also in other, related areas much data is accumulated. For example, it has recently become possible to measure the (relative) activity of thousands of genes at once, using microarrays (sometimes called DNA chips).

It is widely recognized that the mere gathering of data is not sufficient and that the (biological) interpretation of this data is of the utmost importance. Unfortunately, the development of methods for interpreting the data is not keeping up with the tempo with which the data is accumulated. It is clear that many types of questions can only be asked by a computational analysis, and computer science has become an integral part of the research involving biological sequences (of DNA, RNA, or proteins). Some examples of activities for which the computer proves to be an essential tool are

- predicting the locations of genes within large DNA sequences
- comparing a newly sequenced piece of DNA with the known sequences in the databases
- grouping together genes which show similar activity (or expression) patterns when measured under several experimental conditions using microarrays.

The research area combining biology and computer science is known under many names including bioinformatics, computational biology, biocomputing, and biomolecular informatics. Conventional computer methods and algorithms have been applied quite successfully in this area, but the often enormous amounts of data to be analyzed and the complexity of biological systems leave many interesting problems beyond the reach of conventional approaches.

The challenging computational problems of bioinformatics provide interesting opportunities for applying methods from the field of artificial intelligence. See, for example, [3] for an overview of the machine learning approach to bioinformatics. Many problems seem to require the proper translation of biological expertise into heuristics which can make the algorithms more efficient and their results more interesting from a biological point of view.

In this paper, the emphasis is on discussing how methods from the field of uncertainty in AI can be relevant for some challenging problems of bioinformatics. Much of the information bioinformatics has to deal with is uncertain for several reasons. For example, when sequencing (part of) a DNA-molecule, not all nucleotides are identified with a high degree of certainty, comparing new sequences with existing sequences typically does not result in perfect, but only in partial matches, and the measurements of activity of genes (using microarrays or similar tools) is often not very reliable.

The rest of this paper is built up as follows. In the next two sections we will provide a little background information on molecular genetics in general and the human genome project in particular. In section 4, we will

briefly discuss some challenging problems for bioinformatics, and indicate in general terms a few areas where the work on uncertainty in AI might be relevant. In section 5 we try to be more concrete and we zoom in on microarrays, a particular promising and powerful data gathering method, and discuss how Bayesian networks can be employed when analyzing the thus obtained data.

It is not our intention to give a complete overview of the challenges for bioinformatics or to mention all possible applications to bioinformatics of methods and techniques from (uncertainty in) AI, but rather to give an impression of the possibilities.

2 A Few Bits of Molecular Genetics

In this section, some background information on molecular genetics is provided. We will just highlight a few essential facts, which are relevant for the rest of the paper. For a thorough treatment of molecular genetics the interested reader is referred to for example [9] or [17]. It should be stressed that this brief treatment of the field cannot do justice to its rich complexity. It is often jokingly said that about the only rule without exceptions in biology is that all rules have exceptions. We will usually try to avoid explicitly mentioning these exceptions, without oversimplifying our account.

Genetic information is encoded in DNA-molecules, which essentially are long chains of four different types of nucleotides. A nucleotide is a molecule consisting of a base, a sugar (deoxiribose, in the case of DNA) and a phosphate molecule. The four DNA-nucleotides are distinguished by their bases, which are adenine (A), cytosine (C), guanine (G), and thymine (T), respectively, and they are typically referred to by the first letter of their bases. In DNA, these four nucleotides are linearly ordered in strands.

Actually, a DNA-molecule contains two (parallel) strands of nucleotides, which are intertwined to form the well-known double helix shape (which is then folded in a complex manner to fit in the cell nucleus). However, the genetic information is already present in a single strand of the DNA-molecules, since the two DNA-strands are complementary in the sense that A in one strand is paired with T in the other strand and an analogous pairing holds between G and C. Such complementary pairs are called base pairs, and the length of a DNA sequence is often expressed in terms of the number of base pairs.

Due to the complementarity of the two strands, a DNA-sequence can be represented as a sequence of characters out of a four-letter alphabet. For example, a DNA-molecule can be represented by something like the following:

```
GTTCTGTCCTCCGCTGACAAAGCTAACATCAAAGCTACCTGGGACAAAAT...
```

The primary function of DNA is to store information, and this information can become active through its transcription into RNA-molecules, which are quite similar to DNA-molecules, except that RNA is single-stranded and consists of the nucleotides A, C, G and U (uracil), where U in RNA essentially plays the role of T in DNA. DNA nucleotides A, C, G, and T are transcribed into their respective complementary RNA-nucleotides U, G, C, and A.

Typically, only relative small parts of (large) DNA-molecules become active in this sense. Therefore, RNA-molecules are usually much smaller than DNA-molecules. RNA can contain from about fifty to thousands of

A	Ala	Alanine	M	Met	Methionine
C	Cys	Cysteine	N	Asn	Asparagine
D	Asp	Aspartic acid	P	Pro	Proline
E	Glu	Glutamic acid	Q	Gln	Glutamine
F	Phe	Phenylalanine	R	Arg	Arginine
G	Gly	Glycine	S	Ser	Serine
H	His	Histidine	T	Thr	Threonine
I	Ile	Isoleucine	V	Val	Valine
K	Lys	Lysine	W	Trp	Tryptophan
L	Leu	Leucine	T	Tyr	Tyrosine

Table 1: The amino acids and their codes.

nucleotides, whereas DNA-molecules can be many millions nucleotides (or base-pairs) long.

In cells with a nucleus (called eukaryotic cells) the process of transcription from DNA to RNA is complicated by the fact that the transcribed DNA region can contain subregions which do not get transcribed into RNA. The subregions which are dismissed during transcription are called introns. The DNA subgregions which actually do get transcribed into RNA are called exons. (Some organism, such as bacteria, are or are composed of cells without a nuleus or any other intracellular compartments. Such cells are called prokaryotes and have a slightly different process of transcribing DNA into RNA than eukaryotic cells. We will not go further into this.)

There are several kinds of RNA-molecules with different functions, but the main function of RNA is the assembling of amino acids in linear chains, thus forming proteins. Proteins can be represented as sequences composed of an alphabet of 20 amino acids. There is a single character code for the amino acids found in proteins, but also a three character code, where the amino acids are roughly represented by the beginnings of their full names. See Table 2 (The amino acids and their codes) for a list of the amino acids and both their codes.

A sequence of three successive RNA nucleotides is called a codon and can encode one of the 20 amino acids found in proteins or the signal to stop translation of RNA into a protein. The so-called genetic code, relating codons to their associated amino acid is summarized in Table 2 (The genetic code).

For example, the codon `UCG` encodes the amino acid serin. Notice that the genetic code shows some redundancy in the sense that different codons can encode the same amino acid. Actually, only Methionine and Tryptophan have single codons associated with them. Especially in the third position of a codon much redundancy occurs. For example, `UCU`, `UCC`, `UCA`, and `UCG` all encode serin. It follows that not all DNA mutations result in the production of a different protein, even if the mutation occurs in a region which gets transcribed into RNA and then translated. It is generally believed that the genetic code has evolved in a way to (more or less) mimimize the effect of mutations in the DNA. See, for example, [6].

In principle, a single RNA-sequence can give rise to different translations, depending on where the translation area, called the reading frame, starts. For example, the sequence "`UUAUAGC`" encodes "leucine stop" if translation starts at the first position (`UUA` → leucine, `UAG` → stop), but it encodes "tyrosine serin" if translation starts at the second position (`UAU` → tyrosine, `AGC` → serin).

		SECOND POSITION					
		U	C	A	G		
	U	Phe	Ser	Tyr	Cys	U	
F	U	Phe	Ser	Tyr	Cys	C	T
I	U	Leu	Ser	stop	stop	A	H
R	U	Leu	Ser	stop	Trp	G	I
S	C	Leu	Pro	His	Arg	U	R
T	C	Leu	Pro	His	Arg	C	D
	C	Leu	Pro	Gln	Arg	A	
P	C	Leu	Pro	Gln	Arg	G	P
O	A	Ile	Thr	Asn	Ser	U	O
S	A	Ile	Thr	Asn	Ser	C	S
I	A	Ile	Thr	Lys	Arg	A	I
T	A	Met/start	Thr	Lys	Arg	G	T
I	G	Val	Ala	Asp	Gly	U	I
O	G	Val	Ala	Asp	Gly	C	O
N	G	Val	Ala	Glu	Gly	A	N
	G	Val	Ala	Glu	Gly	G	

Table 2: The genetic code.

In most organisms, the reading frames for different proteins do not overlap, but in some viruses the possibility of overlapping reading frames is used heavily to compactify the genetic information. The start of the reading frame is usually encoded by the start codon `AUG`, which encodes the amino acid methionine after the translation has started. This does not mean that every occurrence of the DNA version of this start codon (`ATG`) indicates the start of a region which can be transcribed into RNA. There is much more structure in the transcription process, including the presence of so-called promoter regions almost immediately preceding transcribed regions, and much of this structure is still unknown.

The main function of a protein is that of a catalyst and it is the complex threedimensional form in which a protein folds which determines which specific chemical reactions a protein is able to facilitate or accelerate. Therefore, in addition to knowing the amino acid sequence of a protein it is often important to have insight in its threedimensional structure. Proteins are quite essential to living organism since they play a role in virtually every process in living cells.

A gene can roughly be defined as a functional piece of DNA, which is first transcribed into RNA and then typically is encoded (translated) into a protein. Usually, the introns, the noncoding subregions, of a gene are considered to be part of the gene, and some people are also inclined the include the promoter region as part of the gene. But in some databases only the transcribed RNA-molecule (without introns and promoter) is stored.

Even though all cells in a multicellular organism contain the same DNA-information and genes (leaving aside a few exceptional cases), they can behave quite differently because in different cells different genes can be expressed, i.e., active in the sense of being transcribed into RNA and translated into proteins, at different levels. This cellular differentiation is used by multicellular organisms to form various tissues and organs.

Also, in a single cell the expression levels of genes can vary to respond in a flexible manner to varying environmental conditions. Some diseases may be caused by the presence of one or several mutated genes, but many

malfunctions, such as some forms of cancer, can be explained in terms of extraordinary (either too low or too high) expression levels of some (otherwise normal) genes in particular cells. Thus, not only the presence or absence of a gene in DNA is relevant, its activity (or expression level) does also play a crucial role.

We end this short introduction to molecular genetics by mentioning that the so-called central dogma of information flow in biology states that information can only flow in the direction from DNA to RNA to protein. This nicely summarizes the relation between the central concepts in molecular biology.

3 The Human Genome Project

The total genetic information of an organism is called a genome. For several organisms, including yeast (*Saccharomyces cerevisiae*), the genome is now fully known, in the sense that the exact nucleotide sequence has been determined. The yeast genome is the first completely sequenced eukaryotic genome and consists of about 12 million base pairs.

Based on the sequenced yeast genome, one predicted the existence of more than 6200 yeast genes, many of which were previously unknown. The possibility to systematically search for genes is one of the main scientific contribution of genome projects. The next logical step is to try to find out the function of these genes, which is still ongoing research even in the case of the yeast genes.

Sequencing long DNA-molecules is not a trivial exercise. By laboratory experiments, one can determine the nucleotide sequence of a copy (or clone) of a relatively short piece of DNA. Luckily, this process can be automated using robotic workstations. The problem of fitting together the obtained small sequences in the proper order (as they appear on the long DNA-molecule) is then still a huge problem which would be nearly impossible to attack without the computer. See also section 4.1.

At this time, the description of the human genome (consisting of about 3 billion base pairs) is not quite finished, but it is expected to be completed soon. (The size of the human genome is impressive when compared to that of yeast, but there exist a single-celled micro-organism, called *Amoebe dubia*, with a genome which is about 200 times bigger than the human genome.) In February 2001, two competing human genome projects almost simultaneously published a working draft of the human genome. A special issue of *Nature* (Feb. 15, 2001) includes the description and some analysis of the sequence generated by the publicly sponsored Human Genome Project [12], while *Science* (Feb. 16, 2001) [18] contains the draft sequence reported by a private company called Celera Genomics.

The working drafts of the human genome are incomplete, since the description of the nucleotide sequence still contain several gaps, leaving about 10 percent of the human genome uncovered. Some chromosomes are better covered than others. For example, (the relatively small) chromosome 22, containing about 24 million nucleotides, has been sequenced already in 1999 up to 11 small gaps.

Moreover, the human genome projects are not only intended to produce the nucleotide sequence of human DNA, but also to discover all human genes and their location on the human chromosomes. This goal is much further away from completion, as witnessed by the following facts: The two competing genome projects do not predict exactly the same set of genes, but

different sets which show little overlap among the newly predicted genes, and they disagree for many of the common genes on the assigned location.

One preliminary conclusion the two projects more or less share is that based on the drafts the existence of about 30 thousand human genes in total can be predicted. This number is considerably lower than some previous predictions (which could be as high as 100 thousand), and this fact was mentioned often in popular media reporting on the publication of the drafts to support the suggestion that humans might not be as complex as previously believed.

In fact, the mere number of genes is probably not a good indicator of the complexity of an organism. Moreover, the prediction of roughly 30 thousand as the number of human genes based on the genome drafts of 2001 is not the final word on this issue. For example, in [10] it is argued that a careful comparison of the two sets of predicted genes reveals that the total number of human genes is likely to be much higher than the predicted 30 thousand.

In addition to sequencing the human genome, mapping the human genes and investigate their functions, it is also a goal of the human genome project(s) to chart the (main) variations in the DNA among human beings. Two randomly picked human beings share at least 99.9 percent of their DNA. Some of the 0.1 percent (3 million nucleotides) variation between two individuals have no apparent affect, but some have great influence on appearance, vulnerability to disease, response to medication, etc.

It should be stressed that having a complete description of the human genome does not mean that the human genetic information is then also completely understood. In fact, it can be argued that then the real work still has to be done. For example, knowing the exact sequence of a gene does not imply that one can determine the function of that gene, or even the 3d-structure of the protein encoded by the gene.

Still, the human genome project is generating a wealth of interesting genetic data, such as the description of many previously unknown genes and insight into the relative locations of different genes (which is important since, for example, often genes located close to each other are expressed simultaneously and have the same or related functions).

4 Some Challenges for Bioinformatics

In this section we briefly mention a few challenging problems in the field of bioinformatics. We certainly do not pretend to give a complete list of challenges, but those we do mention are important illustrative examples. We also briefly point out the possible relevance of formalisms, methods and techniques from (uncertainty in) AI. One particular challenging problem, namely the analysis of data from microarray experiments is treated separately in the next section.

4.1 Sequence alignment

If one has sequenced some DNA or protein fragment and wants to know whether this sequence is new or it has been discovered and described before, then one can search the sequences in the (public) databases. It makes little sense to try to find an exact match for the newly found sequence, since usually it is determined rather arbitrary where to start and end sequencing a piece of DNA, some errors might have occured during sequencing, etc. Moreover, it makes sense from a biological point of view to look for similar sequences, since these are likely to be biologically related. Similar sequences

might have similar functions, or they can indicate a close evolutionary kinship.

Let us say that one wants to compare the sequence `TACGATGCTAC` with the following sequences:

```
CTTACGCATGCTAC
TGCTACGAGCTAGT
TGCGATGCTACCGT
TAGTACGTTGCTAC
```

One reasonably good match, or alignment, is given by the sesond sequence in the following way:

```
   TACGATGCTAC
TGCTACGA GCTAGT
```

This match is perfect, except for a single deletion of "T" and a mutation of "C" to "G". To measure the goodness of fit of an alignment and find the best alignments one needs to assign weights to the different possible mutations between nucleotides, and to the deletions and insertions of nucleotides, and then determine how much is minimally required to align the two sequences. The more appropriate the specification of the weights of the different mutations, insertions and deletions is from a biological point of view, the more likely it is that the best alignments are biologically interesting. Sophisticated alignment methods take into account that the different possible mutations are not all equally likely or crucial (since due to the redundancy in the genetic code they need not affect the translation).

Several so-called alignment tools, typically based on dynamic programming, are available. More information on existing alignment methods can, amongst others, be found in [7] or [6].

Perhaps the best known alignment tool is the BLAST (Basic Local Alignment Search Tool) package of [1], which is being improved until today. Although the alignment tools of today can still be improved, perhaps the most useful contribution of AI could be the development of intelligent interfaces for using such tools. There are different versions of BLAST for different tasks (comparing either a nucleotide or amino acid query sequence against either a nucleotide or protein sequence database), and one can set different parameters, such as the length of gaps to be discounted (in order to eliminate the effect of introns in genes). As these alignments tools get more complex, the need will increase for intelligent interfaces, setting appropriate default values for parameters, and providing guidance to users when changing these default values to better fit their specific applications.

An interesting area where there is still room for improving alignment tools is taking account of the uncertainty in the sequences to be aligned. When a sequence is experimentally determined usually not all elements in the sequence are determined with equal certainty, and in the process of aligning multiple partly overlapping sequences one has to propose a so-called consensus sequence which smooths out the misalignments in the previously aligned fragments.

These types of uncertainty have not remained unnoticed. There is even an extended alphabet for referring to disjuntions of nucleotides. For example, "M" denotes "A or C", "V" denotes "A or C or G", etc. Sometimes probability distributions are used to represent the certainty of a particular nucleotide being present at a particular location in a sequence. However, these representations are not fully exploited in the present alignment tools.

It might also be interesting to develop and exploit a notion of "fuzzy alignment". There has been some, but not much, work using fuzzy techniques in this area. See [14].

Perhaps the main challenge related to sequence alignment is the alignment of more than two sequences. The efficiency of the dynamic programming approaches breaks down in this case. Multiple alignment is an important issue, especially for the problem of sequencing large DNA-molecules (as in the genome projects), where one usually experimentally obtains relatively short and partly overlapping subsequences which then have to be pieced together similar to completing a giant puzzle.

Solving this puzzle is not only difficult because the pieces are not known with complete certainty, but also facts like the existence of regions with many repeating sequences in the human genome complicate matters. For example, it is difficult to determine which of the following three sequences should be made out of the pieces `CAGC`, `GCAT`, `ATAT`, and `ATCG`:

```
CAGCATATATATCG
CAGCATATATCG
CAGCATATCG
```

Simply choosing the minimal sequence incorporating all small sequences may be biologically incorrect. As is often the case in bioinformatics, in multiple alignment there is a need for improving brute force methods by applying smart algorithms which employ general biological knowledge and possibly specific information about the small sequences, for example the lenght of the overlap between these sequences, or the average number of small sequences covering a single position.

4.2 Gene prediction

Not all of the DNA information appears to have a clear function like encoding proteins. In fact, it is estimated that about 90 percent of the human genome is noncoding, and this type of DNA is usually called "junk DNA". (Some people object to this name since of course it is perfectly possible that this kind of DNA does have a function, although perhaps less obvious than the genes.)

It is a challenging problem to efficiently find (or predict) in a long DNA-sequence the coding regions (or genes). To tackle this problem one of course needs to make use of the available biological information, This information is typically of a heuristic nature such as the following. In the human genome one has observed that around the start regions of many (but not all) genes one can find so-called CpG islands, i.e., regions with a relatively high frequency of the occurrence of the sequence `CG` (often written as CpG to distinguish it from a C-G pair where the two nucleotides appear in complementary DNA-strands). See, for example, [7] for an application of Markov chains to this problem of gene prediction.

As mentioned before, the RNA transcript of a gene is often not obtained by continuous transcription of the gene, but it involves so-called RNA splicing where some unwanted (noncoding) subregions of the gene, called introns, are dismissed and the transcripts of the remaining coding subregions (exons) are then joined together. Finding the exon/intron boundaries (splice junctions) of genes is perhaps an even more challenging problem than finding the genes.

Within a gene, exons can be quite sparsely distributed. Cases are known of exons with a total length of a few thousand nucleotides which are spread

out over a gene which is more than a million nucleotides long. Also here, the main challenge is to extract informative heuristics out of the relevant biological knowledge, such as the following. It is known that introns almost always start with `GT` and end with `AG`, but these sequences frequently occur without indicating the start or the end of an intron.

As witnessed by the fact that the two human genome projects predict quite different sets of genes, predicting human genes is still a nontrivial problem. But even imperfect predictions can be quite useful. Typically, results from bioinformatics (obtained in what is called the "dry lab") need experimental verification (in what is called the "wet lab"), but as long as computer predictions can sufficiently focus biological experiments such that the chance of success is sufficiently high, then bioinformatics is worth the effort.

It should be mentioned that the problem of gene prediction will remain important even after the completion of the human gene project, since for many types of experiments humans cannot be used and one has to use other (more or less genetically related) organisms (such as mice), for which the genome might not always be known. (In fact, the mouse genome is comparable in size to the human genome.) More generally, one can say that improving the prediction of coding regions is an important step towards the real understanding of genetic information. Perhaps machine learning approaches will be able to infer from known coding regions of sequences what are the (main) characteristics of coding regions, thereby providing valuable biological insights.

4.3 Protein structure prediction

As mentioned before, the function of a protein is determined by its 3d-structure, which is typically quite complex. One can envision that in the future proteins will be classified by their 3d-structures which in turn will be associated with different (types of) functions. Unfortunately, for many (in fact, most) proteins the 3d-structure is presently unknown.

Although the three-dimensional structure of a protein is determined by its amino acid sequence, it can presently not be accurately predicted from this sequence. This prediction problem is much harder than the gene prediction problem, and much less heuristic information is available to exploit. Therefore, in this case, the machine learning approaches which might help to obtain the necessary information, can perhaps be relatively more important. Oherwise, similar remarks apply to this problem as in the cans of the gene prediction problem.

4.4 Distributed data sources

The number of (often huge) biological database is quite impressive, and the available information grows at an enormous rate. To get an impression of the amount of data, one can for example have a look at [3], where several (more than 100) of the most important of these databases are listed: 22 under the heading "Databases over databases", 17 "Major public sequence databases", and 65 "Specialized databases".

Sometimes, these databases store complementary information and are connected through hyperlinks. For example, after finding a gene in GENBANK you can follow a link to find information on its associated protein in SWISSPROT. However, it also happens that the same kind of service

(say gene finding) is available from different sites which can use different implementations, often giving different results.

The task of manually extracting all (or most) relevant information is quite tedious and in some cases (when one wants information about many sequences) almost impossible. This problem is likely to get much worse in the future, since the available information keeps growing.

The automatic extraction of information from distributed, partly complementary, partly conflicting, data sources is an important research area for AI. Although this problem is relevant for many applications (for example, weather prediction, where information from different satellite instruments have to be combined), improving data mining from distributed biological databases seems especially interesting since it clearly addresses a need and the possible benefits are quite substantial. See [2].

Some subproblems of combining distributed data sources, such as evidence combination and dealing with inconsistent information, have been studied extensively in the field of uncertainty in AI. Also, since some of the relevant information about the sequences can be found in the form of free text in online literature, methods for text mining might be relevant.

4.5 Genetic networks

The cases where a particular high-level biological function can be associated with (the expression of) a single gene are rare exceptions. Usually, such a function is performed by a network (or pathway) of genes working together and influencing (either positively or negatively) each others expression levels.

It is important to get insight into these genetic networks in order to really understand the associated biological functions and to be able to influence these functions, for example to cure diseases. Proper medication should correct the effects of extraordinary (either too low or too high) concentration of proteins, which are closely related to the expression levels of their associated genes. Ideally, medication should not have any side effects and not disturb other processes than one it targets. In the remote future, it might even be possible to use protein level understanding of diseases and their possible cures to produce personalized medication which individually minimizes side effects.

Experiments can be performed to provide pieces of information concerning genetic networks. For example, one can look at consequences of knocking out (i.e., making inactive) or overexpressing (i.e., making more active than usual) a single gene. The results of these experiments can often only be stated in qualitative, or semi-quantitative, terms. ("If gene1 is knocked out, then gene2 is expressed much more and gene3 slightly less".) These experimental results should then be combined with general background knowledge. ("Gene3 is probably not influenced directly by gene1".)

In [19] a (purely qualitative) method of inferring genetic networks based on abductive inference is discussed. In general, the problem of inferring genetic networks is likely to benefit from methods for combining qualitative and (semi-)quantitative information, which have been studied within the field of uncertainty in AI.

5 Microarray Analysis

A particular promising and powerful data gathering method uses so-called DNA microarrays (sometimes also called DNA chips). For details on this

technique and its potential the reader is referred to [13]. There are several variants of DNA microarrays, with different applications, but the most widely used benefit of the technique of DNA microarrays is that it allows one to simultaneously compare the expression levels of many genes (in some cases, all the genes in a genome) under different conditions. For example, one can compare a cancerous cell with a normal cell, or sample the temporal evolution of expression levels of genes is a cell after it has received a certain stimulus.

Microarray technology is still relatively new and under development. Several authors warn that the results of microarrays are not always reliable or useful. See [11] for a description of some of the dangers, and see [5] for an argument that one needs a careful experimental design of microarray experiments to ensure that these experiments will help accumulate knowledge and not just enormous amounts of useless data. There is also a need for standardizing the publication of microarray data. See [4].

A typical microarray experiment results in an image of (many) colored dots, where each dot is associated with a particular gene and its color is related to the expression level of that gene (in comparison to the expression level in a reference cell). For example, a red dot means that the expression level of the associated gene is relatively high and a green dot means that this expression level is relatively low (and an intermediate color means something "in between").

Repeating the experiment under different experimental conditions or at different times after some stimulus results for each gene in a list of (relative) expression levels. It is then standard practice to cluster genes with similar expression levels. The underlying idea is that co-regulated genes (i.e., genes with similar expression levels) probably have related functions. Many clustering techniques have been developed, but of course all require a wisely chosen distance or similarity measure to identify similar expression levels.

Although these clustering techniques may be useful for finding co-regulated genes, it is difficult to use them for the discovery of gene interactions (or genetic networks). This problem is addressed using Bayesian networks in [8]. Here some recent advances in the area of learning (partial) Bayesian networks are used to infer Bayesian networks which are (most) likely to represent gene interactions. The authors admit that learning Bayesian networks is difficult (and at the moment still infeasible) for the huge number of genes for which microarrays can measure the expression levels. However, they apply their method to a subset of 800 of the 6177 yeast genes for which expression data is reported in [16].

It is interesting to note that the Bayesian networks approach of [8] is not proposed as an alternative to clustering methods, but rather as a tool which can complement more traditional clustering methods. The clustering analysis of [16] is used to single out a subset of interesting genes (and an even smaller subset of 250 genes in 8 clusters to perform some of the robustness analysis). The fact that the Bayesian networks approach can induce some structure (within clusters of genes) is interesting from a biological point of view, although it remains to be studied how the induced (probabilistic) structures relate to causal genetic networks.

In [15] some further limitations of the clustering methods for analyzing gene expression levels are pointed out. For example, these methods measure similarities that exist over all of the measurements, while obscuring relationships existing over only a subset of the data. Another example is that other types of information, such as clinical data, cannot easily be combined with the similarity measure of expression levels. The authors propose

an alternative approach based on the language of probabilistic relational models (which extend Bayesian networks to a relational setting).

6 Conclusion

Automatic information processing and computer analysis are essential to handle and understand the huge amounts of biological data which is becoming available today. Bioinformatics is likely to be of crucial importance for the next decades in both fundamental and applied biological and medical research. Even with the increasing computing power of modern computers, many tasks in bioinformatics are only feasible if one employs intelligent algorithms which make use of biological expertise.

Since much of the biological data is uncertain, incomplete or otherwise imperfect, and typically the relevant biological knowledge is of a heuristic nature, it is clear that many problems in bioinformatics have several aspects in common with problems studied in the field of uncertainty in AI.

We listed several of challenges for bioinformatics where formalisms, techniques and methods from uncertainty in AI can be relevant. These challenges range from rather basic and tradional sequence alignment to the analysis of recently developed microarray techniques.

References

[1] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, and D.J. Lipman, Basic local alignment search tool, *Journal of Molecular Biology* 207 (1990) pp. 403–410.

[2] C. Andorf, D. Caragea, J. Reinosi-Castillo, A.Silvescu, V. Honavar, and D. Dos, Ontology-driven information extraction and knowledge acquisition from heterogeneous, distributed, autonomous biological data sources, *Proceedings of the IJCAI-2001 Workshop Knowledge Discovery from Distributed, Heterogeneous, Dynamic, Autonomous Data Sources*, Seattle, Washington (2001) pp. 1–12.

[3] P. Baldi and S. Brunak, *Bioinformatics: The Machine Learning Approach*, MIT Press, Cambridge MA (1998).

[4] A. Brazma et al., Minimum information about a microarray experiment (MIAME) - toward standards for microarray data, *Nature Genetics* 29 (2001) pp 365–371.

[5] G. A. Churchill and B. Oliver, Sex, flies and microarrays, *Nature Genetics* 29 (2001) pp 355–356.

[6] P. Clote and R. Backofen, *Computational Molecular Biology: An Introduction*, Wiley, Chichester (2000).

[7] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison, *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, Cambridge U.P. (1998).

[8] N. Friedman, M. Linial, I. Nachman, and D. Pe'er, Using Bayesian networks to analyze expression data, *RECOMB 2000*, Tokyo (2000) pp. 127–135.

[9] A.J.F. Griffiths, J.H. Miller, D.T. Suzuki, R.C. Lewontin, W.M. Gelbart, *An Introduction to Genetic Analysis (seventh edition)*, Freeman, New York (1999).

[10] J.B. Hogenesch et al., A comparison of the Celera and Ensembl gene sets reveals little overlap in novel genes. *Cell* 106 (2001) pp. 413–415.

[11] J. Knight, When the chips are down, *Nature* 410 (2001) pp. 860–861.

[12] E.S. Lander et al. (International Human Genome Consortium), Initial sequencing and analysis of the human genome, *Nature* 409 (2001) pp. 860–921.

[13] B. Phimister (ed.), The Chipping Forecast, *Nature Genetics Supplement*, vol 21, no. 1 (1999).

[14] K. Sadegh-Zadeh, Fuzzy genomes, *Artificial Intelligence in Medicine* 18 (2000) pp. 1–28.

[15] E. Segal, B. Taskar, A. Gasch, N. Friedman, and D. Koller, Rich probabilistic models for gene expression, *Bioinformatics* 1 (2001) pp. 1–10.

[16] P.T. Spellman et al., Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization, *Molecular Biology of the Cell* 9 (1998) pp. 3273–3297.

[17] T. Strachan and A.P. Read, *Human Molecular Genetics (2nd edition)*, BIOS Scientific Publishers, Oxford (1999).

[18] J.C. Venter et al., The sequence of the human genome, *Science* 291 (2001) pp. 1304–1351

[19] B. Zupan, I. Bratko, J. Demsar, J.R. Beck, A. Kuspa and G. Shaulsky, Abductive inference of genetic networks, *AIME01: Biennial Conference of the European Society for Artificial Intelligence in Medicine*, Cascais, Portugal (2001).