GENOMES, MAPS AND SEQUENCES

MICHAEL S. WATERMAN*

Departments of Mathematics and Molecular Biology
University of Southern California
Los Angeles, CA 90089-1113

INTRODUCTION

In the second half of the twentieth century biology has progressed at breakneck speed. James Watson and Francis Crick in 1953 proposed the now famous double helical structure for DNA. This structure gave a physical model for how one DNA molecule can divide and become two identical molecules. On this point they wrote one of the most famous sentences of science: "It has not escaped our notice that the specific pairing we have postulated immediately suggests a possibly copying mechanism for the genetic material." And that copying mechanism based on the adenine (A)–thymine (T) and guanine (G)–cytosine (C) base pairing turned out to be correct and is the foundation of molecular genetics. While about 100 years earlier Mendel gave an abstract model of inheritance, Watson and Crick gave a specific molecular model that can be studied and manipulated. The last 50 years of molecular biology has been in large part based on the Watson-Crick discovery. See Lewin (1990) and Alberts *et al.* (1983) for excellent general accounts of the subject.

There are two other macromolecules that must be mentioned. Proteins provide the structural molecules and enzymes of which organisms are built. DNA was not initially thought to be the molecule of inheritance as it is composed of only four components (the bases mentioned above). Proteins with their twenty amino acids seemed far more likely to hold the complex secrets of inheritance. While it was known experimentally before Watson and Crick that DNA was the basis of inheritance, their model flushed the question of the genetic code into the open. How does DNA encode the information for proteins? Since DNA is a sequence of bases or nucleotides, a sequence n long has 4^n possibilities and the seeming simplicity of DNA vanishes. For example, a sequence of length 1000 has over 10^{600} possibilities while there are only 10^{80} elementary particles in the universe; a sequence of length 133 has about 10^{80} possibilities. While some mathematically clever coding schemes were proposed, nature has chosen a simple three bases per amino acid code, where the triplets specifying successive amino acids in a protein appear sequentially and non-overlapping along the DNA sequence.

The third important macromolecule is RNA, another nucleic acid of four bases. RNA has several roles. One is messenger-RNA (mRNA), a RNA complementary copy of the DNA or gene corresponding to a protein. The mRNA is read by the ribosome, a

* This research supported by grants from the National Institutes of Health (GM-36230) and the National Science Foundation (DMS-90-05833)

complex of protein and structural RNA molecules, and is translated into a sequence of amino acids defining a protein. These structural RNA molecules in the ribosome are known as rRNA. Originally thought to be less central, RNA has assumed an increasingly important role in the last decade. Some RNA molecules have been shown to have enzymatic activity. There is evidence for example that the scores of proteins in the ribosome are not essential to its activity and that the three structural rRNAs might be able to translate mRNA into protein unassisted. This lends support to an evolution of life from RNA molecules, a point of view called "the RNA world."

The genetic code was worked out in the 1960s and protein and nucleic acid sequences began to be read. In those early days RNA was more easily sequenced than DNA and proteins more easily than nucleic acids, but all sequencing was very difficult. Then about 1976 Maxam and Gilbert at Harvard and Sanger at Cambridge proposed two methods that accelerated DNA sequencing by two orders of magnitude. Almost immediately exciting and unexpected discoveries were made. One of the first was the so-called intron-exon nature of eukaryotic genes that we will now describe. In prokaryotic organisms, those without a nucleus, the gene encoding a protein is an uninterrupted sequence of triplets (called codons). *E. coli*, a prokaryote that lives in our gut, had become the model organism for molecular biology because it is easily grown and manipulated. Imagine the surprise when it was discovered that eukaryotic genes were interrupted by non-coding DNA, called intervening sequences or introns. The coding intervals are called exons (for expressed). This discovery has several implications of interest.

First of all, why would an organism evolve a mechanism such as an intron? At first glance it seems to be hopelessly inefficient and complex. However, if they were at a selective disadvantage, introns should disappear. One suggestion is that the exon units can be more easily recombined into new proteins than the corresponding events would occur in an uninterrupted gene. Another is that the intron-exon structure of genes is primitive and that only in the prokaryotic lineage have introns disappeared. The truth is that no one is certain as to why introns exist.

Secondly, forgetting why introns exist, another question is how we get from the intron-exon gene to the mRNA to be translated into protein. The answer lies in a mechanism known as splicing. The DNA is translated into RNA, then the introns are cut out, and the exons are spliced together to make the mRNA. This splicing mechanism has been well characterized in the last few years. (See Fig. 1.)

A third point about introns brings us to a topic in computational molecular biology. Even in prokaryotic organisms, genes are not entirely trivial to recognize. Three triplets are stop codons signaling the end of the amino acid chain so one technique is to find longer stretches with no stops; in addition there is a standard codon ATG found at the beginning of genes. Even so, mistakes can be made and statistical methods have been devised. Allowing introns to interrupt a coding sequence of 900 bases, say, and lengthen the gene into 10,000 or more bases greatly complicates the scientist's problem of recognizing genes in DNA sequences. Since we can and are sequencing DNA more and more rapidly, this is a central problem of much practical importance. The most successful

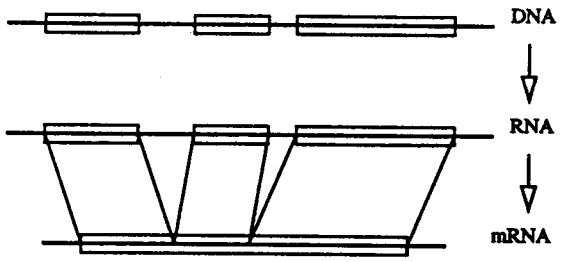

Figure 1: DNA, RNA, and mRNA

method to date combines several imperfect gene prediction methods via a neural network. Those who desire prediction methods based directly on biological models must search for an even deeper understanding of the splicing problem.

Nucleic acid sequence data has been collected into international data bases since 1982. (See Fickett & Burks (1989) for a survey of DNA databases.) The three major databases EMBL (Europe), GenBank (USA) and DDBJ (Japan) are for practical purposes identical today although certain distinctions existed in earlier years. The database content measured in nucleotides approximately doubles every two years. Figure 2 shows this growth. While sequences over 100,000 bases exist, the median sequence length is about 1000 bases. Of course, there are an increasing number of properties of the sequences that are of scientific content and only a few of the most important properties such as gene locations can be found in the databases.

The mathematical discussions in this paper will be organized into three sections.

1. Mapping DNA. Maps are representations of landmarks on sequences and are consequently less informative than sequences. They are easier to construct and are very useful. We will briefly mention genetic maps and will present physical mapping in somewhat more detail.

2. Comparing Sequences. Once sequences are obtained they are compared with themselves and with other sequences. There is a series of related comparison problems and solutions and in addition corresponding map comparison problems.

3. Genomes. Finally we will take a quick look at entire genomes—all the DNA of an organism—and speculate about the problems of the future.

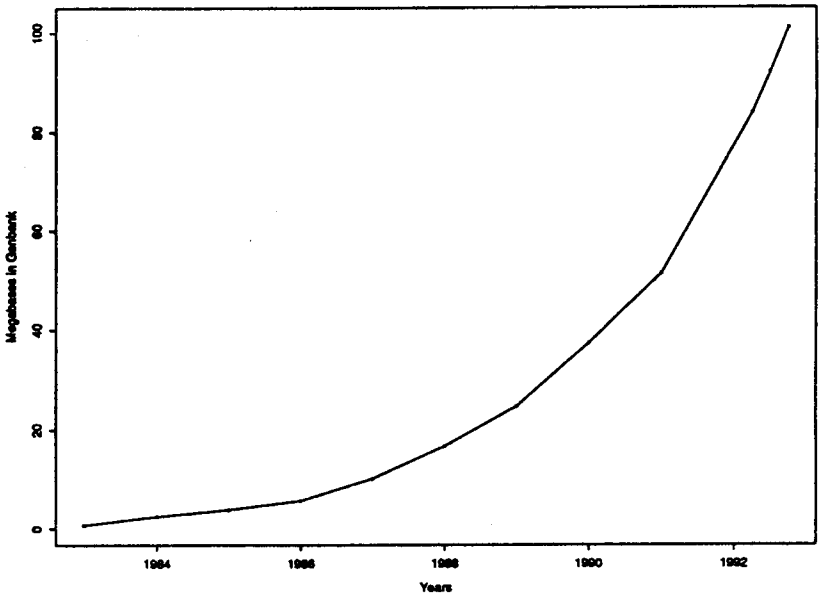

Figure 2: DNA Database growth.

2. MAPPING DNA

2.1 Genetic Mapping

The idea for chromosome cartography or mapping is based on the idea that genes are contained in and linearly arranged along (the DNA of) a chromosome. Thomas Hunt Morgan made central contributions to this area, utilizing the large chromosomes of *Drosophila melanogaster* in his research. Sturtevant, who was a student of Morgan, constructed in 1913 the first genetic map of 6 genes or traits. The map gave the approximate locations of these 6 traits with different recombination probabilities or distances between them. In *Drosophila* there are a number of single genes with mutations causing observable traits such as curly wings, white eyes, or stubbly bristles that can

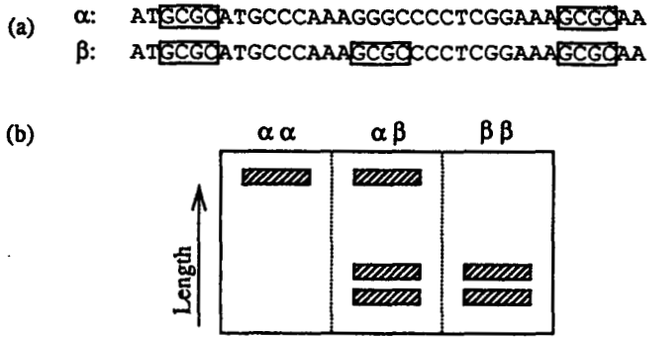

Figure 3: (a) Two DNA RFLP's; (b) Three genotypes on a Southern blot.

be mapped. What is important is that a single chromosome location can be associated with an observable trait in the organism.

Genetic mapping in the post Watson-Crick era is undergoing massive growth. This is due to the 1980 suggestion of Botstein *et al.* (1980) that slight variations in DNA sequence between a pair of homologous chromosomes could provide such markers for humans and other eukaryotes. The pair of homologous chromosomes refer to the pair of chromosomes, one from the mother and one from the father. Two unrelated humans differ in about 1 base per 1000 so such variation between homologous chromosomes is expected. Secondly there are developed for use a few hundred restriction enzymes that cut double stranded DNA at short specific locations on the double helix. Variations in sequence implies variation in the distance between cutting sites or restriction sites since cutting sites can appear or disappear because of variation in the DNA sequence. Since these fragment lengths can be measured, this gives a large number of chromosome locations and observable traits without the necessity of finding single genes with mutations that result in observable traits. The restriction fragment length polymorphisms are called RFLPs. An ultimate goal of mapping these variations is to then determine the approximate location of various disease genes. Recently molecular geneticists have approximately located genes with mutations that result in Huntington's disease, cystic fibrosis, polycystic kidney disease and others. The genes for Huntington's disease and cystic fibrosis have both subsequently been cloned and sequenced. The book by Ott (1991) is a general reference to genetic mapping in human chromosomes.

To return to the Botstein idea, variation in DNA can cause restriction sites present in one sequence to be absent in another. In Figure 3 the restriction enzyme under consideration is HhaI that cuts at the sequence GCGC. Our Figure 3 is modeled after

Lander(1989). Notice that the complementary strand of the double helix, read from right to left, is also GCGC. The feature holds for almost all these patterns. In Figure 3(a) the top chromosomal sequence α GCGC occurs twice while in the bottom sequence β it occurs three times. Therefore the top sequence will have one restriction fragment while the bottom has two. They can be visualized in a Southern blot where patterns for $\alpha\alpha$, $\alpha\beta$ and $\beta\beta$ are shown.

Returning to our mapping problem, there is an unknown probability θ of a recombination between the RFLP and another trait locus. The maximum likelihood estimate $\hat{\theta}$ is used as a measure of linkage between the loci. The probability $\theta = 1/2$ means the loci are unlinked while $\theta = 0$ means they are at essentially the same chromosome location with no recombination possible between them.

Modern genetic mapping is not restricted to RFLPs. Other genetic variation or polymorphisms can be used. One such polymorphism comes from a variable number of tandem repeats of a fairly short sequence (VNTRs). These VNTRs are more polymorphic than RFLPs and can distinguish the DNA of a parent from their child's. Other repetitive sequences, the minisatellites, have variable numbers of repeats and can be used for mapping. For example there are many CA repeats. All of these loci have their own set of experimental and analytical positive and negative features, creating a very active area of research.

2.2 Physical Mapping

In genetic mapping, the goal is to locate genes or loci on the chromosome where the distance between them is the recombination distance. Now we turn to more direct measurements of the distance between loci, in particular where the distance is measured in number of nucleotides. Our initial problems will arise from the ability to cut DNA with restriction enzymes and to measure the length of the resulting restriction fragments, as discussed in Section 2.1. The goal is to obtain the map of the order and location of the restriction enzyme sites along the DNA molecule. In Figure 4 maps are shown for two enzymes α and β, and α and β together. There are 3 possible maps, α alone, β alone, and α and β together. The physical distance between the sites is proportional to number of nucleotides.

Some nice graph theory is associated with these maps. Interval graph theory originated with the biologist Benzer (1959) who was studying the structure of bacterial genes. While every schoolchild today knows a gene is a linear word over a four letter alphabet, Benzer's work was basic to deciding that fact. He had experimental data on the overlap of pairs of fragments of the gene and he showed the data consistent with linearity, founding a new area of discrete mathematics. The corresponding data for restriction maps is knowledge about whether or not intervals between restriction sites overlap or not.

When the digest goes to completion, that is the enzyme cuts at all sites, we obtain all intervals between adjacent sites. The intervals are arbitrarily indexed in Figure 5.

Overlap data can be summarized in incidence matrices, $I(\alpha, \beta) = (x_{ij})$, where

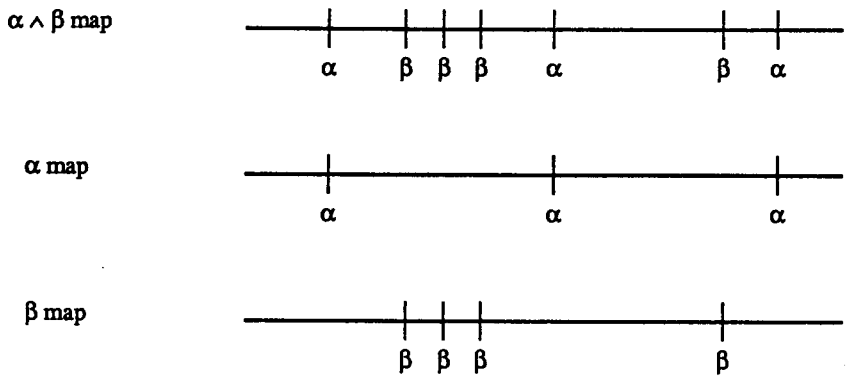

Figure 4: The three possible restriction maps from two enzymes α and β.

$$x_{ij} = \begin{cases} 1 & \text{if } \alpha\text{−fragment } (i) \cap \beta\text{−fragment } (j) \neq \emptyset \\ 0 & \text{if the intersection is } \emptyset\,. \end{cases}$$

It is elementary to show

$$I(\alpha,\beta) = I(\alpha, \alpha\wedge\beta) I^t(\beta, \alpha\wedge\beta)\,.$$

How do we know that $I(\alpha,\beta)$ is consistent with a restriction map and how do we find that map from $I(\alpha,\beta)$? For our problem,

$$I(\alpha,\beta) = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 1 \end{pmatrix}.$$

By rearranging rows and columns, we obtain the following staircase shape for the matrix

$$\begin{array}{c} \\ 3 \\ 4 \\ 2 \\ 1 \end{array} \begin{array}{c} \begin{array}{ccccc} 4 & 1 & 3 & 5 & 2 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \end{array}.$$

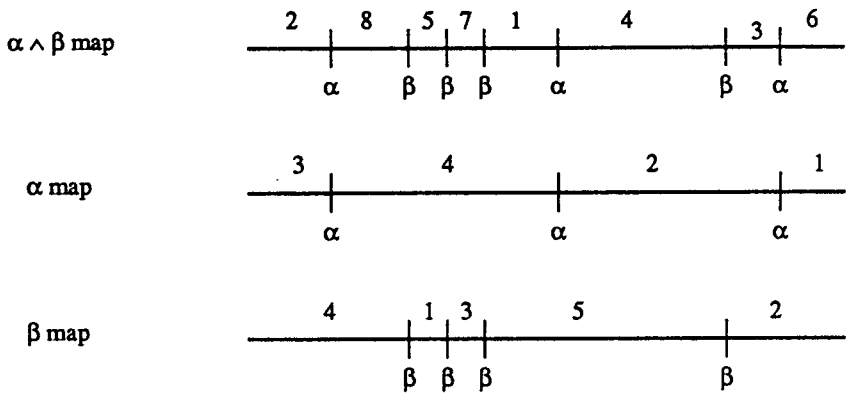

Figure 5: Restriction maps from two enzymes α and β with labelled intervals.

This staircase property of the permuted incidence matrix is a characterization of bipartite interval graphs with no isolated vertices. Griggs amd Waterman (1986) apply the ideas and results of interval graphs to restriction maps.

There is a concept and experimental practice of partial digestion. In these experiments a site is cut with probability $p \in (0,1)$, not $p = 1$ as in complete digestion. This raises the possibility of intervals, such as $\overline{5-7-1}$ in the $\alpha\wedge\beta$ map, composed of adjacent single digest intervals. For the α,β overlap graph we introduce two such intervals in addition to the complete digest intervals.

$$I^* = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ \overline{4-2} \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \quad \overline{3-5-2} \\ \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix} \end{array}$$

A row permutation, this matrix becomes

$$\begin{array}{c} \\ 1 \\ 2 \\ \overline{4-2} \\ 4 \\ 3 \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 & \overline{3-5-2} \end{array} \\ \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix} \end{array}$$

Notice that in the columns 1's occur uninterrupted by 0's. This **consecutive ones property for columns** characterizes **interval graphs**: a graph whose vertices can be put into 1-1 correspondence with a set of intervals of real numbers whose vertices are connected if their corresponding intervals have non-empty intersection. This of course is a version of the original problem of Benzer and was established by Lekkerkerker & Boland (1962). Linear time algorithms for obtaining and testing for interval graphs can be found in Booth and Leuker (1976).

2.2.1 The Double Digest Problem

With Southern blots as symbolized in Figure 3, we obtain length measurements of all single and double digest fragments. Often these measurements are only approximate. For the example of this section, let x_1, x_2, x_3, x_4 denote the lengths of the 4 α-digest fragments, $y_1, y_2 \ldots y_5$ the length of the 5 β-digest fragments, and $z_1, z_2, \ldots z_8$ the lengths of the $\alpha \wedge \beta$ digest. Given these lengths, the **double digest problem** DDP is to find the maps consistent with the data. Nathans and Smith (1975) introduced the idea of constructing restriction maps from length data. Generally we have

$$A = \{x_1, \ldots, x_n\} ,$$
$$B = \{y_1, \ldots, y_m\} ,$$

and

$$A \wedge B = \{z_1 \cdots z_l\}$$

If there are no coincident cut sites, $l = n+m-1$. Taking the ideal case of no measurement errors,

$$\sum_{i=1}^{n} x_i = \sum_{i=1}^{m} y_i = \sum_{i=1}^{l} z_i .$$

We have not made precise our criteria for a good solution. Most approaches to this problem fall into two categories. The first is what we call the travelling salesman or permutation approach. In this setting the task is to find permutations $\sigma \in S_n$ and $\mu \in S_m$ so that (σ, μ) specifies a map. Set

$$S = \left\{ s : s = \sum_{i=1}^{r} a_{\sigma(i)} \text{or} \quad \sum_{i=1}^{t} b_{\mu(i)} : \quad 1 \le r \le n, \quad 1 \le t \le m \right\}$$

Index S so that

$$S = \{s_j : \quad 0 \le j \le l\} ,$$

$s_j \le s_{j+1}$, and $s_0 \equiv 0$. The double digest implied by S is

$$D(\sigma, u) = \{z_i(\sigma, \mu) = \quad s_j - s_{j-1} \quad \text{for} \quad j \in [1, l]\}$$

The quality of $D(\sigma, \mu)$ is measured by how near it is to the real double digest data, $\|D(\sigma, \mu) - A \wedge B\|$. In Goldstein & Waterman (1987), $\|\cdot\|$ was defined by

$$\|D(\sigma, \mu) - A \wedge B\| = \sum_i \frac{(z_i(\sigma, \mu) - z_i)^2}{z_i}$$

Various approaches have been taken to solve the problem. Pearson (1982) simply looked at all $n!m!$ permutations of the single digests. Goldstein and Waterman proposed a simulated annealing algorithm. Any heuristic approach to the traveling salesman problem should be adaptable to this problem.

Another approach to DDP is the set partition approach. That is for the α digest, for example, partition the l double digest lengths into n disjoint classes:

$$\begin{array}{cccc} z_{1,1} & \cdots & z_{1,n_1} \\ z_{2,1} & \cdots & z_{2,n_1} \\ & \cdots & \\ z_{n,1} & \cdots & z_{n,n_n} \end{array}$$

and check the fit by

$$\sum_{i=1}^{n} \| \sum_{j=1}^{n_j} z_{i_s j} - x_i \| .$$

Fitch *et al.* (1983) proposed a solution that took essentially this approach.

As shown in Goldstein & Waterman (1987), it is relatively straightforward to show DDP is in the class of NP complete problems conjectured to have no polynomial time solution. Garey & Johnson (1979) is a standard reference to NP complete problems. More surprisingly, if we lay down restriction sites according to a Poisson process, it can be proved that there is an exponentially increasing number of exact solutions as the length increases, with probability one. Only one (or two if we consider left/right symmetry) can be biologically correct. Therefore it is hard to find a solution which in turn is unlikely to be that in which the biologist is interested. Biologists cope with this problem by staying safely on this side of asymptotics. Schmitt & Waterman (1991) look at the multiple solutions more closely, and a complete solution to characterization of these multiplicities is given in Pevzner(1994).

2.2.2 Partial Digest Problems

We have mentioned partial digestion above in the introduction to Section 2.2. The procedure for partial digest mapping is experimentally demanding but mathematically trivial.

The DNA is end labelled, so that when the lengths are measured we "see" only pieces that have the label. In the example shown in Fig. 6 with the α partial digest we would only see the fragments shown. Measuring these four lengths, a map is easily constructed for the enzyme.

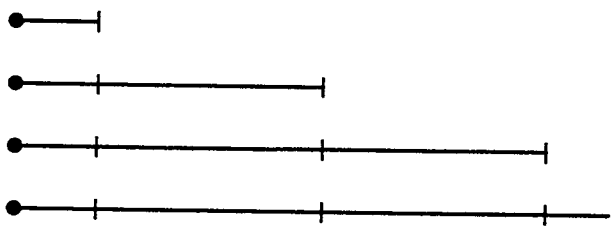

Figure 6: A Partial Digest.

When the DNA is not end labeled, the mathematical complexity changes greatly. This new partial digest problem (PDP) can be stated as follows. Given a set $\mathcal{L} = \{x_1, \ldots, x_n\}$ of lengths, the problem is to find distinct points on the line $c_1 < c_2 < \ldots$ such that

$$\mathcal{L} = \bigcup_{1 \le i < j} \{c_i - c_j\} .$$

There is a beautiful algorithm for PDP based on Rosenblatt & Seymour (1982) and presented by Lenke & Werman (1992) and by Skiena *et al.* (1990). In Naor (1992) the algorithm, which computes all possible generating functions for possible maps, is shown to be pseudo-polynomial. That is the time complexity is bounded by a polynomial function of the size of the problem, n, as well as the largest fragment length. Naor also presents a backtracking algorithm more appropriate to real data with multiplicities and errors.

Naor also studies a related problem, the probed partial digest problem, PPDP. In this problem, an interval is labeled and all partial digest intervals containing that interval are measured. The formal statement is that the PPDP length set $\mathcal{L}$ is solved by a set of $p+q$ distinct points on a line $x_1 < x_2 < \cdots < x_p < y_1 \cdots < y_q$ where

$$\mathcal{L} = \bigcup_{1 \le i \le p} \{y_j - x_i\} .$$

Newberry & Naor (1992) give a lower bound on the multiplicity of solutions to this problem.

3. COMPARING SEQUENCES

Actually it is profitable to look at a DNA sequence as the finest scale map it is possible to make of DNA. As we suggested in the introduction, the total database of sequenced DNA is approximately doubling every two years, and with genome projects gaining momentum, this rapid increase will continue for some time. In this section we will survey a few aspects of nucleic acid and protein sequences. First of all, since obtaining sequence is so important, we first outline the simplest and often used technique of shotgun sequencing. Then we survey the area of sequence comparisons.

3.1 Sequencing DNA

Most currently employed techniques for sequencing DNA rely on the ability to routinely read a small segment of DNA sequence of length l, where $l \in [300, 500]$ bases, with reasonable accuracy. Essentially, this is accomplished by dividing the DNA into four samples and treating each sample with an enzyme that cuts after a specific letter. Then the samples are run side by side in a gel and the sequence can be read from the gel. Let us assume this technology is available.

The technique we now describe for obtaining sequences longer than l is known as shotgun sequencing. The idea is to break the longer sequence of length L into random fragments of length l. That is any interval I, $|I| = l$ has the likelihood $(L - l + 1)^{-1}$. Then these fragments are read sequentially. The longer sequence is determined by the ability of the experimenter to find overlaps of the sequenced fragments and to put the sequence together like a jigsaw puzzle. See Figure 7 for a schematic of this process.

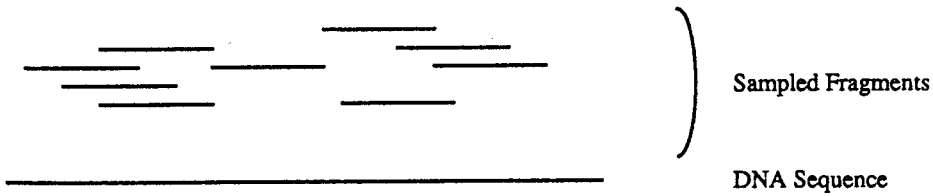

Figure 7: Sequencing fragments

Obviously, there must be a minimum believable overlap, t. $t = 1$ for example would have overlap "seen" every $1/(2 \times p)$ comparisons where $p = P$ (Two random bases are identical). With independent, identically, and uniformly distributed letters, $p = 1/4$. We need to set t so that we have a small chance ϵ of calling overlap. Suppose n= the number of fragments. Then the chance of spurious overlap is estimated by

$$P \text{ (spurious overlap)} = P(\bigcup_{i,j} \{\text{fragment } i \text{ and fragment } j \text{ overlap by } \geq t\}$$
$$\leq 2\binom{n}{2}\sum_{k\leq t} p^k \leq 2\binom{n}{2}\frac{p^t}{1-p}.$$

Solving

$$2\binom{n}{2}\frac{p^t}{1-p} = \epsilon,$$
$$t = \log_{1/p}\left(\frac{n(n-1)}{\epsilon(1-p)}\right).$$

Therefore if $n = 1000$, $\epsilon = 10^{-2}$, and $p = 1/4$, we have $t = 13$. Since there is more data than just pairwise overlaps in sequence assembly, the t used in practice is often as small as 7 or 8.

Another analytic question of much importance is how large n must be to obtain 95% of the underlying sequence of length L. Some formulas developed by Lander & Waterman (1988) for the progress of physical mapping experiments are relevant here. First set the notation

L = length of DNA to be sequenced
l = length of sequenced fragments
n = number of sequenced fragments
α = n/L
t = amount of overlap needed to detect overlap
θ = t/l
σ = $1-\theta$
c = redundancy of coverage = nl/L.

Several assumptions underlie the formulas given below. We assume $L \cong L - l + 1$ so that, for example $n/L \cong$ probability of a base starting a new fragment. We also assume all overlaps $\geq t$ are detected perfectly, with no errors.

As sequencing proceeds, the fragments fall into apparent islands of one or more members based on their sequence. Of course some actual islands go undetected because of small overlap. Islands are often called "contigs" and the gaps between islands are called oceans. All claims below are approximate for n large.

I. The expected number of apparent islands is $ne^{-c\sigma}$.

II. The expected number of apparent islands with $j \geq 1$ fragments is

$$ne^{2c\sigma}(1-e^{-c\sigma})^{j-1}.$$

III. The expected number of apparent islands of at least two fragments is

$$ne^{-c\sigma} - ne^{-2c\sigma}.$$

IV. The expected number of fragments in an apparent island is $e^{c\sigma}$.

V. The expected length of an apparent island is

$$L\left[\frac{e^{c\sigma}-1}{c} + 1 - \sigma\right].$$

VI. The probability an ocean of length at least kl follows an island is $e^{-c(k+\theta)}$.

3.2 Sequence Comparison

Given a sequenced DNA, both RNA and protein sequences can be inferred. One way to understand the function and evolution of sequences is to compare a sequence to all sequences already collected in the databases. It can be enormously useful to know that the sequence has a unusual, non-random similarity to a sequence already in the database. Evolution maintains useful sequences over vast evolutionary time so that sequence similarity can be suggestive of the original biological function of the sequences. For example, before the defective gene for cystic fibrosis was finally cloned (Riordan *et al.*, 1989) and sequenced there was no hypothesis about the structure or function of the gene. After sequencing the gene, the inferred protein sequence was compared to the protein sequence database, and a family of membrane transport proteins were similar. This gave an immediate hypotheses as to the structure and function of the gene that turned out to be extremely valuable in guiding the next experiments.

How then are sequence comparisons performed? While there are a wide range of methods, here we will present a family of dynamic programming methods for sequencing comparison. The following discussion based on Waterman (1984) and Waterman (1989). The dynamic programing methods are quite simple to write down and they give rigorous answers to these questions. Let us first look at the problem of comparison, often called alignment. Take two sequences $\mathbf{x} = x_1x_2\ldots x_n$ and $\mathbf{y} = y_1y_2\ldots y_m$. We compare the sequences by aligning them, writing the x-sequences over the y-sequence. Two aligned letters $\frac{x}{y}$ denote the identity as in $\frac{A}{A}$ or a substitution as in $\frac{A}{G}$. To complicate things, letters can be inserted or deleted (an indel), denoted by $\frac{A}{-}$. This means an A was inserted into the x-sequence or deleted from the y-sequence. Alignments can be enumerated by identifying the "matched" letters. If there are a total of k identities and substitutions. then there are $\binom{n}{k}\binom{m}{k}$ corresponding alignments. Summing over k

$$\sum_k \binom{n}{k}\binom{m}{k} = \binom{n+m}{n},$$

indicating a huge number of alignments for sequences of length 1000.

The idea of optimal alignment is to assign a score to each alignment. The simplest scoring scheme is to give $s(x,y)$ to each matched $\begin{matrix} x \\ y \end{matrix}$ and to give $-\delta$ for each deletion $\begin{matrix} x \\ - \end{matrix}$ or $\begin{matrix} - \\ y \end{matrix}$. Usually $s(x,x) > 0$ and $s(x,y) < 0$ for at least some x, y pairs. Other more complex scoring schemes will be discussed later.

3.3 Global alignment

This section continues the above discussion: how to find optimal alignments of $\mathbf{x} = x_1 x_2 \cdots x_n$ and $\mathbf{y} = y_1 y_2 \ldots y_m$ with scoring function $s(x,y)$, where $s(x,-) = s(-,y) = -\delta$. Dynamic programming provides an elegant way to find these optimal alignments. Set

$$S(\mathbf{x}, \mathbf{y}) = \max\{\text{score}: \text{ all alignments of } \mathbf{x} \textit{ and } \mathbf{y}\} .$$

For our algorithm,

$$S_{i,j} = S(x_1 \ldots x_i, y_1 \ldots y_j)$$

and $S_{0,j} = S(\emptyset, y_1 \ldots y_j) = -\delta \times j$, $S_{i,0} = S(x_1 \ldots x_i, \phi) = -\delta \times i$, $S_{0,0} = S(\emptyset, \emptyset) = 0$. The algorithm is based on the three ways an alignment can end:

$$\begin{matrix} \ldots x_i \\ \ldots y_j \end{matrix} \quad \text{or} \quad \begin{matrix} \ldots x_i \\ \ldots - \end{matrix} \quad \text{or} \quad \begin{matrix} \ldots - \\ \ldots y_j \end{matrix} .$$

The optimal score $S_{i,j}$ must correspond to alignments that end in one of these three ways. The sequence prior to the ending must be optimally aligned. Therefore

$$S_{i,j} = \max\{S_{i-1,j-1} + s(x_i, y_j),\ S_{i-1,j} - \delta,\ S_{i,j-1} - \delta\} .$$

The cases where i and/or j equal 1 are handled by the boundary values for $S_{k,l}$ when $k \cdot l = 0$.

	$\emptyset$	G	G	T	G	A	A	A	G	G	C
$\emptyset$	0	−2	−4	−6	−8	−10	−12	−14	−16	−18	−20
T	−2	−1	−3	−3	−5	−7	−9	−11	−13	−15	−17
A	−4	−3	−2	−4	−4	−4	−6	−8	−10	−12	−14
A	−6	−5	−4	−3	−5	−3	−3	−5	−7	−9	−11
T	−8	−7	−6	−3	−4	−5	−4	−4	−6	−8	−10
T	−10	−9	−8	−5	−4	−5	−6	−5	−5	−7	−9
T	−12	−11	−10	−7	−6	−5	−6	−7	−6	−6	−8
G	−14	−11	−10	−9	−6	−7	−6	−7	−6	−5	−7
T	−16	−13	−12	−9	−8	−7	−8	−7	−8	−7	−6
G	−18	−15	−12	−11	−8	−9	−8	−9	−6	−7	−8
G	−20	−17	−14	−13	−10	−9	−10	−9	−8	−5	−7

Table 1: Table comparison matrix

Figure 8 shows the schematic (or cartoon as molecular biologists put it) of an alignment. An example is now given where x =TAATTTGTGG y =GGTGAAAGGC and $s(x,y) = \begin{cases} 1, & x = y \\ -1, & x \neq y \end{cases}; -\delta = -2.$

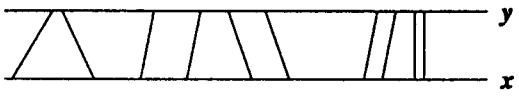

Figure 8: Global alignment

```
T A A T T T G T G G -
      |       | |
- G G T G A A A G G C

T A A T T T G T G G -
      |       | |
G - G T G A A A G G C

T A A T T T G T G G -
      |       | |
G G - T G A A A G G C
```

Table 2: Global alignments

The time cost of the algorithm is clearly $O(n^2)$. The optimal alignments are obtained by starting at S_{nm} and asking what $S_{i,j}$ lead to $S_{n,m}$. Each such $S_{i,j}$ is one part of an alignment. This so-called backtracking scheme recursively gives all optimal alignments. For our example the optimal alignments are in Table 2.

3.4 Fitting in

In many examples a pattern or sequence of interest will be known and the problem is to fit the pattern into a larger sequence. That is, find those contiguous subsequences of $\mathbf{y} = y_1 \dots y_m$ where $\mathbf{x} = x_1 \dots x_n$ is well aligned. Formally

$$F(\mathbf{x},\mathbf{y}) = \max\{S(\mathbf{x}, y_k \dots y_l) : 1 \le k \le l \le m\} .$$

This problem too is easily solved by the above recursion with a subtle but essential modification. Set $F_{0,0} = 0$, $F_{0,j} = 0$ and $F_{i,0} = -i\delta$. Then, as before, compute

$$F_{i,j} = \max\{F_{i-1,j-1} + s(x_i, y_j),\ F_{i-1,j} - \delta,\ F_{i,j-1} - \delta\} .$$

The score $F(\mathbf{x},\mathbf{y})$ is obtained by

$$F(\mathbf{x},\mathbf{y}) = \max\{F_{n,j} : \quad 1 \le j \le m\} .$$

The schematic for this problem is in Figure 9.

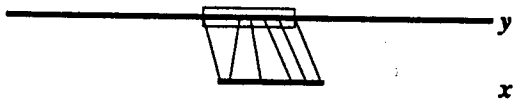

Figure 9: Fitting in alignment.

The logic behind this recursion is of course closely related to that for global alignment. Deletion of the initial part of the y sequence should have 0 cost; that is accomplished by $F_{0,j} = 0$. Deletion of the final part of the y sequence also should have 0 cost; this is accomplished by equation 3 maximizing over $F_{n,j}$, which deletes $y_{j+1} \cdots y_m$ at 0 cost.

3.5 Local alignment

We now come to the most useful algorithm for sequence studies, the so-called local algorithm. When a new sequence is compared to a database of sequences, there is no clear idea of what might be found. The entire search sequence x might be very close to a sequence in the database, so that we should compute $S(\mathbf{x}, \mathbf{y})$. The entire search sequence might be within a database sequence, so that we should compute $F(\mathbf{x}, \mathbf{y})$. The search sequence might contain an entire database sequence so that we should compute $F(\mathbf{y}, \mathbf{x})$. Or even more likely, a portion of the sequence, x will have an unusual resemblance to

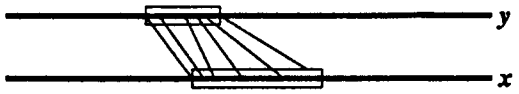

Figure 10: Local alignment

a portion of y and none of the above solutions are correct. The schematic for this new problem is in Figure 10 where the boxes indicate the portions of x and y to be aligned.

Formally define

$$H(\mathbf{x}, \mathbf{y}) = \max\{S(x_i \dots x_j, y_k \dots y_l) : 1 \le i \le j \le n,\ 1 \le k \le l \le m\} .$$

Again a delicate modification of the basic recursion will do the trick. Set $H_{i,j} = 0$ if $i \cdot j = 0$. Then define

$$H_{j,l} = \max\{S(x_i \cdots x_j, y_k \cdots y_l) : 1 \le i \le j \quad \text{and} \quad 1 \le k \le l\} .$$

The modified recursion is

$$H_{i,j} = \max\{H_{i-1,j-1} + s(x_i, y_j), H_{i-1,j} - \delta, H_{i,j-1} - \delta, 0\} .$$

The original problem is solved by proving that

$$H(\mathbf{x},\mathbf{y}) = \max\{H_{i,j} : 1 \le i \le n,\ 1 \le j \le m\} .$$

The new subtlety is the introduction of "0." This acts to eliminate any alignment that has negative score, allowing an alignment to begin anywhere with 0 cost for "end deletion."

The example treated earlier is now recalculated in Table 3 for the local algorithm. The best local alignment is found by tracing back from $H_{9,4} = 3$ resulting in an identical GTG from each sequence.

	G	*G*	*T*	*G*	*A*	*A*	*A*	*G*	*G*	*C*
T	0	0	1	0	0	0	0	0	0	0
A	0	0	0	0	1	1	1	0	0	0
A	0	0	0	0	1	2	2	0	0	0
T	0	0	1	0	0	0	1	1	0	0
T	0	0	1	0	0	0	0	0	0	0
T	0	0	1	0	0	0	0	0	0	0
G	1	1	0	2	0	0	0	1	1	0
T	0	0	2	0	1	0	0	0	0	0
G	1	1	0	3	1	0	0	1	1	0
G	1	2	0	1	2	0	0	1	2	0

Table 3: Local alignment example

3.8 Indel costs

In biology, long stretches of sequence can be inserted or deleted in one evolutionary step. It does not seem correct to weight a deletion of k letters by $-k\delta$. Instead we might weight such a deletion by $-w(k)$, where $w(k)$ is an arbitrary non-negative function. The global alignment algorithm becomes

$$S_{i,j} = \max\{S_{i-1,j-1} + s(x_i, y_j), \max\{S_{i-k,j} - w(k) : 1 \le k \le i, \\ \max\{S_{i,j-k} - w(k) : 1 \le k \le j\}\} .$$

The time cost escalates from $O(n^2)$ to $O(n^3)$.

When $w(k) = \alpha + \beta k$, it is possible to reduce the cost to $O(n^2)$ again by running 3 recursions instead of 1 as for the original algorithm. See Gotoh (1982) for this algorithm. It is possible to argue that slowly increasing functions, such as $w(k) = a + b\log(k)$ are even more appropriate. For these concave indel functions more complex algorithms running in time $O(n^2 \log n)$ are presented in Miller & Myers (1988) and Galil & Giancarlo (1989).

4. GENOMES

The word genome is defined as the collection of DNA contained in an organism. In the case of humans it means the DNA in the 23 linear chromosomes, about 3×10^9 nucleotides. In *E. coli* the genome is one circular chromosome of about 4.7×10^6 nucleotides. Recently "genome analysis" has become a hot topic with the advent of the Human Genome Project. We will discuss the U.S. project but it is worldwide, coordinated by HUGO. Roughly speaking the Project's goal is to characterize the human genome, although it is actually much broader than that. The publicity of the project has created much scientific interest in the general area, including so-called informatics that has been defined to mean anything computational or mathematical relating to biological information. Below Section 4.1 discusses the Human Genome Project in more detail, Section 4.2 sketches a recent topic in molecular sequences and human evolution, and Section 4.3 concludes with a sampling of the new challenges that the Human Genome Project brings to the mathematical sciences.

4.1 The Human Genome Project

The usual description of the broad goal of the HGP is to acquire basic information needed to further basic scientific understanding of human genetics and of the role of various genes in health and disease. This is a goal that can easily be described to a politician or to a non-scientist. It is our belief that the goals are much deeper and broader. To understand human biology it is necessary and much easier to study the biology of a wide range of organisms. This is basically due to the conservative nature of evolution. Important molecules are conserved over vast amounts of evolutionary time. Plants have a version of the hemoglobin molecule, as do bacteria which have been on this planet for 3.5–4×10^9 years. Obviously these molecules have evolutionary descendents that are utilized for new purposes but much can be learned from these comparisons. Of course it is much easier to grow bacteria than mammals to mention only practical and

not ethical considerations. Two humans differ by about 1 base in 1000; we are 99.9% similar. A human and a chimpanzee differ by about 2 bases in 100; we are 98% similar. Lab animals, especially the laboratory mouse, are used for experiments vital to human health. It would be a waste of money and time not to have parallel genome projects for other organisms. These organisms are called "model organisms" in the literature of the HGP and include *E. coli* as well as other bacteria, the yeast *S. cerevisiae*, the fruitfly *Drosophila melangaster*, the worm *C. elegans*, and the mouse. The collection of map and sequence information from these and other organisms will be the foundation of biology for the next century.

Let me now sketch some of the goals of the first five years of the project as they appear in a joint U.S. National Institutes of Health and U. S. Department of Energy publication, **Understanding Our Genetic Inheritance: The U.S. Human Genome Project.** Of course the outline below reflects that of the publication.

4.1.1 Mapping and sequencing of the human genome

It is estimated that there are 50,000 to 100,000 human genes, but this number is not well estimated. Some 2,000 of these genes have been mapped. Since there is variation in human sequence, the question of "whose sequence?" often arises. Actually "the" initial sequence will be a composite of many human genomes. Population variations are important and can be pursued as well; they just are not the initial goal.

Genetic maps were briefly discussed above. Genetic maps are sequences of markers separated by estimated genetic dictators. The unit of distance is a centimorgan, after the geneticist T.H. Morgan. A centimorgan roughly corresponds to 10^6 bases of DNA. The 5 year goal is to provide a fully connected genetic map with markers an average of 2 to 5 centimorgans apart. Each marker is to be identified by a short DNA sequence that is unique in the human genome, a sequence tagged site or STS.

Recall that physical maps are sequences of sites with the intersite distances measured in physical length, such as number of bases. There are many variants of physical maps. One 5 year goal is to produce STS maps of all chromosomes with the markers at approximately 100,000 base intervals. Another type of physical map is an overlapping clone map. DNA can be inserted into various types of clones: λ clones of 15,000 bases, cosmids of 34,000 bases, BACS (bacterial artifical chromosomes) of 10^5 bases, or YACS (yeast artificial chromosomes) of 10^5 to 10^6 bases. The 5 year goal is to cover large parts of the human genome with overlapping clone maps.

The physical maps of overlapping clones for the smaller size clones is a useful preparation for DNA sequencing. The 5 year sequencing goal is to sequence 10^7 bases of human DNA in large continuous stretches. This goal is to help develop and validate technology. In addition, the 5 year goal is to improve existing and develop new methods of DNA sequencing that will speed up sequencing and reduce the cost from \$1 per base to \$.50 per base.

4.1.2 Model organisms

Some reasons for studying model organisms were given above. The mouse is an extremely important model organism for the HGP and the 5 year goal is to construct a genetic map of the mouse genome. In addition 2×10^7 bases of DNA from a variety of model organisms is to be sequenced, on sequences that are 10^6 bases long, in the course of improving existing and developing new methods of DNA sequencing.

4.1.3 Informatics: data collection and analysis

Here the goal is to develop effective software and database designs for support of large mapping and sequencing projects. Key to the subject of the present paper is the development of algorithms and analytic tools to interpret genomic information.

Other goals relate to ethical, legal and social considerations, research training, technology development and transfer.

Notice that the emphasis is on the data collection aspects of the HGP and scant attention is given to the understanding of the data. This is misleading of course. As soon as the cystic fibrosis gene is cloned and sequenced the race is off to understand the gene and its defect, and to devise treatments including gene therapy. This is a general rule: produce the information base and science will proceed at a rapid pace.

4.2 Mitochondrial Eve

Evolution has been discussed briefly at several places in this paper. In this section we look at a recent controversy in human evolution resulting from the analysis of DNA sequences. The root question of the time and place of human origins as well as mechanisms of evolution is of great interest. Molecular data hold great promise, but as the following discussion shows, progress is made with great difficulty. Yet it is just these questions that must be addressed with the flood of molecular sequence data. These are serious challenges for mathematical biology.

The late Allan Wilson and colleagues used mitochondrial (mt)DNA to address evolutionary questions. mtDNA is useful because it accumulates substitutions rapidly and because it is matriarchal, passed from mother to offspring, and can be used to trace ancestry without the recombination occurring in the lineage of most genes. In Cann *et al.* (1987) they argued that the most recent common ancestor of all mtDNA sequences in human populations today lived in Africa around 200,000 years ago. The paper locates mitochondrial Eve in time and place and made a huge impression on the scientific and popular press.

How was such a wonderful conclusion reached? First of all, the data was the presence or absence of 195 restriction sites from 147 individuals who were Africans, Asians, Caucasian, native Australians, and New Guineans. Secondly the data was analyzed by the method of parsimony. In parsimony analysis, the goal is to find the evolutionary tree of minimum length, where length is the minimum number of evolutionary changes along the tree to explain the existing data. A central problem is that at about 15 species or individuals all possible trees cannot be exhaustively searched by brute force.

At 147 individuals, no known method can rigorously search all trees. Heuristics must be employed as the problems are NP hard.

Cann *et al.* (1987) found a tree with 312 evolutionary changes. The first branch had all Africans on one branch and a mixture of African and non-Africans on the other. From this an African origin was argued. That takes care of place. Time was estimated by assuming a uniform rate of change in mtDNA. The statistical issues here have not been satisfactorily resolved.

Since the paper appeared several studies have cast doubt on the analysis. See Goldman & Bartony for an excellent overview (1992). It turns out that there are at least 10,000 trees with only 307 changes, most of which do not indicate an African origin. In addition the parsimony analysis does not always give the best estimate of the true tree. In some well-defined situations, the method of maximum likelihood is superior. A model of evolution along the branches of the trees is specified and the maximum likelihood tree is to be chosen. Here the computational requirements are even more severe than in the case of parsimony.

4.3 Challenges of the Future

The last section was meant to illustrate the new problems motivated by the HGP and by modern molecular biology in general. Evolutionary questions arise everywhere: Discover the evolutionary history of genes, species, groups within species, ... There is no one model that is adequate. For example, the simple binary, branching tree might be made completely invalid by transfer of genetic material between species that **cannot** interbreed. Mathematical models of evolution led biology early in this century. Today we are behind the experimentalist who has data and questions far beyond our knowledge and power to answer.

Sequences occupied much of this paper. As databases grow exponentially and the comparison questions grow rapidly as well, we must devise entirely new ways to rapidly compare strings. It may even be that analog will replace digital computation in this area. Certainly people in the field of combinatorial pattern matching will be kept busy by these new problems.

Of course there are genome maps of every flavor being produced. How to compare maps is an active area. Huang & Waterman (1992) look at extensions of dynamic programming sequence comparison algorithms to physical map comparisons. Sankoff & Goldstein (1989) take on a new problem when they look at minimum rearrangements of gene orders to change one chromosome map into another.

This paper has not discussed one of the most important and difficult problems in theoretical biology, that of predicting 3-dimensional protein structure from the linear sequence of amino acids. See Creighton (1989) for example. No clean solution has been proposed thus far and this field will remain very active as DNA sequences along with the protein sequences they encode continue to appear.

There are important and fascinating new problems in this area. Generally they arise in connection with specific biological data and questions. It is our job to translate the

biology into mathematics and computer science to obtain relevant problems for our subject as well as for biology.

REFERENCES

Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K., and Watson, J. D.. (1983). *Molecular Biology of the Cell*, Garland Publishing Inc., New York & London.

Benzer, S. (1959). On the topology of genetic fine structure, *Proc. Natn. Acad. Sci.* **45**, 1607–1620.

Botstein, D., White, R. L, Skolnick, M. H., and Davis, R. W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms, *Am. J. Hum. Genet.*, **32**, 314–331.

Cann, R. L., Stoneking, M, and Wilson, A. C. (1987). Mitocondrial DNA and human evolution, *Nature*, **325**, 31–36.

Creighton, T. . (Ed.) (1989). *Protein Structure: A Practical Approach.* IRL Press, Oxford, New York, Tokyo.

Fickett, J.W. and Burks, C. (1989). Development of a database for nucleotide sequences. In *Mathematical Methods for DNA Sequences* (Ed. Waterman, M. S.), 1–34.

Fitch, W.M., Smith, T.F., and Ralph, W.W. (1983). Mapping the order of DNA restriction fragments, *Gene*, **22**, 19–29.

Galil, Z. and Giancarlo, R. (1989). Speeding up dynamic programming with applications to molecular biology. *Theor. Comput. Sci.*, **64**, 107–118.

Garey, M.R. and Johnson, D.S. (1979). *Computers and Intractability: A Guide to the Theory of NP-Completeness.* Freeman, San Francisco.

Goldman, N. and Barton, N.H. (1992). Genetics and Geography, *Nature*, **357**, 440–441.

Goldstein, L. and Waterman, M.S. (1987). Mapping DNA by stochastic relaxation. *Adv. Appl. Math.*, **8**, 194–207.

Gotoh, O. (1982). An improved algorithm for matching biological sequences. *J. Mol. Biol.*, **162**, 705–708.

Huang, X. and Waterman, M. S. (1992). Dynamic programming algorithms for restriction map comparison, *CABIOS* **8**, 511–520.

Lander, E.S. and Waterman, M.S. (1988) Genomic maping by fingerprinting randm clones: a mathematical analysis. *Genomics*, **2**, 231–239.

Lander, E.S. Analysis with restriction enzymes, In *Mathematical Methods for DNA Sequences* (Ed. Waterman, M. S.), 35–52.

Lekkerkerker, G.C. and Boland J.C. (1962). Representation of a finite graph by a set of intervals on the real time. *Fund. Math.*, **51**, 45–64.

Lemke, P. and Werman, M. On the complexity of inverting the autotocorrelation function of a finite integer sequence, and the problem of locating n points on a line given the $\binom{n}{2}$ unlabeled distances between them, Manuscript.

Lewin, B. (1990). *Genes IV*, Oxford University Press, Oxford, New York, Tokyo, Melbourne.

Miller, W., and Myers, E.W. (1988). Sequence comparison with concave weighting functions. *Bull. Math. Biol.* **50**, 97–120.

Naor, D. (1992). Mapping algorithms for DNA partial digestion; a survey. Manuscript.

Nathans, D. and Smith, H.O. (1975) Restriction endonucleases in the analysis and restructuring of DNA molecules, *Ann. Rev. Biochem.*, **44**, 273–293.

Newberg, L. and Naor, D. (1992). A lower bound on the number of solutions to the probed partial digertion problem. *Adv. Appl. Math* To appear.

Oh, J. (1991). *Analysis of the Human Genetic Linkage*, The Johns Hopkins Univ. Press, Baltimore & London.

Pearson, W. (1982). Automatic construction of restriction maps, *Nucleic Acids Res.*,**10**, 217–227.

Pevzner, P.A. (1994). DNA physical mapping and alternating Eulerian cycles in colored graphs. In press. *Algorithmica.*

Rosenblatt, J. and Seymour, P.D. (1982). The structure of homometric sets, *Siam. J. Alg. Disc. Math.*, **3**, 343–350.

Riordan, J.R. *et al.* (1989). Identification of the Cystic Fibrosisgene: cloning and characterization of complementary DNA. *Science*, **245**, 1066–1072.

Sankoff, D. and Goldstein, M. (1989). Probabilistic models of genome shuffling. *Bull. Math. Biol.*, **51**, 117–124.

Schmitt, W. and Waterman, M.S. (1991). Multiple solutions of DNA restriction mapping problems. *Adv. Appl. Math.*, **12**, 412–427.

Skiena, S. S., Smith, W. D., and Lemke, P. (1990). Reconstructing sets from interpoint disstances (Extended abstract), Proc. of the 6th Ann. Symp. on Computational Geometry, ACM Press, 332–339.

Understanding Our Genetic Inheritance. The U.S. Human Genome Project: The First Five Years. FY 1991–1995. U.S. Department of Health and Human Services. and U.S. Department of Energy joint publication DOE/ER-0452P. Available from the National Technical Information Service, U.S. Department of Commerce, Springfield, VA 22161.

Waterman, M. S. and Griggs, J. R. (1986). Intervals graphs and maps of DNA. *Bull. Math. Biol.*, **48**, 189–195.

Waterman, M. S. (1984). General methods of sequence comparison. *Bull. Math. Biol.*, **46**, 473–500.

Waterman, M. S. (1989). Sequence alignments. In Waterman (Ed.) *Mathematical Methods for DNA Sequences*, 53–92.

Waterman , M. S.(Ed). (1989). *Mathematical Methods for DNA Sequences*, CRC Press, Boca Raton, Florida.

CELL PROTRUSIONS

GEORGE OSTER

Departments of Molecular and Cellular Biology and Entomology
University of California, Berkeley, CA 94720

ALAN S. PERELSON

Theoretical Division Los Alamos National Laboratory
Los Alamos, NM 87545

INTRODUCTION

How do cells move? There is no shortage of theories, but a definitive answer is still elusive. Here we will present some models for cell motions along with proposals for experiments to address the theories. We will restrict ourselves to the problem of cell protrusion, although the models apply to other motility phenomena as well.

Different protrusion phenomena extend at different characteristic rates. Figure 1 shows data for protrusion rates of lamellipodia, filopodia and the acrosomal process of *Thyone*. Their velocities vary dramatically, which suggests that the force driving them may originate from different physical mechanisms. We will present here models for each of these processes.

LAMELLIPOD PROTRUSION

Lamellipodia are broad, flat cytoplasmic protrusions that spread out in front of a moving cell. Experiments have shown that there is surface flow of cytoplasm rearward from the leading edge as the lamellipodium extends forward [20]. Conservation of mass ensures that there must be a central flow of cytoplasm forward to provide the material for extension. The question we address is: what forces drive the extension of the leading lamella? Amoebae pseudopodia bear superficial resemblance to lamellipodia, but there is evidence that they may operate by a different mechanism, and so we will not deal with them here; models of the "frontal contraction" and "cortical tractor" hypotheses for amoeboid motion can be found in [16, 17, 25, 26, 27].

An elegant set of models have been developed by Dembo, Alt, and their coworkers based on the idea that cytoplasm can be modeled as a "contractile fluid" [1, 7, 8, 9]. They model the cytogel as a two-phase, viscous fluid (i.e. a Navier-Stokes system) with a term added that permits the gel phase