CONF-9307199--Extd.Abstr.
FG03-93ER61638

EXTENDED ABSTRACTS AND TOPICS FOR DISCUSSION

CONF-

OPEN PROBLEMS OF
COMPUTATIONAL MOLECULAR BIOLOGY
third international workshop

Telluride Summer Research Center
Telluride, CO

July 11 - 25, 1993

ORGANIZERS: Andrzej K. Konopka and Peter Salamon
COORDDINATOR: Danielle A.M. Konings
SPONSORS: U.S. Department of Energy - Human Genome Program
CONVEX Computer Corporation
BioLingua-44 Research Cons.

CALENDAR OF EVENTS

*Open Problems of Computational Molecular Biology **3***

Telluride, CO, July 11 - 25, 1993

	MONDAY	TUESDAY	WEDNESDAY	THURSDAY	FRIDAY	SATURDAY	SUNDAY
	07/12/93	07/13/93	07/14/93	07/15/93	07/16/93	07/17/93	07/18/93
AM 10:00	**Konopka 41**	**Rosen 61**	**Lawrence 45**	**Argos 1**	**Modena 54**	General	**Schuster 63**
AM 10:30	discussion	**Rosen**	**Lawrence**	**Argos**	**Modena**	discussion	**Schuster**
AM 11:00	**Hamming 17**	discussion	discussion	discussion	discussion	*DNA*	discussion
AM 11:30	**Hamming**					*sequence*	
NOON 12:00	discussion					*analysis*	
PM 5:30			**Hamming's**	**Rosen's**	**Casti's**		
PM 6:00			author	author	author		
PM 6:30			evening	evening	evening		
PM 7:00							
PM 7:30							
PM 9:00	**Casti 11**	**Peuzner 59**			General	**Björklund**	**Stadler 64**
PM 9:30	**Casti**	**Peuzner**			discussion	**Gordon 3**	**Stadler**
PM 10:00	discussion	discussion			*inform.*	discussion	discussion
PM 10:30					*theories*		
PM 11:00					*and biology*		
PM 11:30							
MIDNIGHT							

	MONDAY	TUESDAY	WEDNESDAY	THURSDAY	FRIDAY	SATURDAY	SUNDAY
	07/19/93	07/20/93	07/21/93	07/22/93	07/23/93	07/24/93	07/25/93
AM 10:00	**Macken 51**	**Fickett 12**	**Hatrick 27**	**Konings 38**	**Salamon**	**Konopka**	General
AM 10:30	**Macken**	**Fickett**	**Hatrick**	**Konings**	**Salamon**	*closing*	discussion
AM 11:00	discussion	discussion	discussion	discussion	discussion	*remarks*	
AM 11:30							
NOON 12:00							
PM 5:30					panel	panel	
PM 6:00					discussion	discussion	
PM 6:30					for	for	
PM 7:00					other	other	
PM 7:30					workshops	workshops	
PM 9:00	**Istrail 36**	**Wootton 65**	**Heringa 35**	discussion			
PM 9:30	**Istrail**	**Wootton**	**Heringa**	*genome*			
PM 10:00	discussion	discussion	discussion	*projects*			
PM 10:30							
PM 11:00							
PM 11:30							
MIDNIGHT							

Numbers after the names indicate pages on which the corresponding abstracts begin (no number means no abstract).

Time and "substance" of events may change.

CONTENT BY AUTHOR's NAME[1]:

[1]Names of authors and co-authors who do not participate in the workshop are printed in *italic*.

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

Patrick Argos

European Molecular Biology Laboratory
Meyerhofstraase 1
Postfach 10.22.09
Heidelberg
Germany

ABSTRACT 1:

PREDICTION OF PROTEIN FOLDING PATHWAYS

Recent nuclear magnetic resonance (n.m.r.) hydrogen exchange experiments on five different proteins have delineated the secondary structures formed in trapped, partially folded intermediates. The early forming structural elements are iden-tifiable with a method devised to predict folding pathways. The technique that the sequential selection of structural fragments such as alpha-helices and beta-strands involved in the folding process is founded upon the maximal burial of solvent accessible surface from both the formation of internal structure and substructure association. The substructural elements are defined objectively by major changes in main-chain direction. The predicted folding pathways are in complete correspondence with the n.m.r. results in that the formed structural fragments found in the folding intermediates are those predicted earliest in the pathways. The technique has also been applied to proteins of known tertiary structure and with fold similar to one of the five proteins examined by n.m.r. observations, suggesting conservation of a secondary structural framework or molten globule about which folding nucleates

ABSTRACT 2:

A NEW METHOD TO CONFIGURE PROTEIN SIDE-CHAINS IN HOMOLOGY MODELLING

Protein homology modelling typically involves the prediction of side-chain conformations in the modelled protein while assuming a main-chain trace taken from a known tertiary structure of a protein with homologous sequence. It is generally believed that the need to examine all possible combinations of side-chain confirmations poses the major obstacle to accurate homology

modelling. Methods proposed heretofore use only discrete or limited searches of the side-chain torsion angle space to mitigate the combinatorial problem and also rely on simplified energy functions for computational speed. The configurational constraints are typically based upon use of frequently observed torsion angles, fixed steps in torsion angles, or oligopeptide segments taken from tertiary structural databanks that are similar in sequence and conformation with the target structure. A more fundamental approach has been explored for several protein structures and it is demonstrated that the combinatorial barrier in side-chain placement hardly exists. Each side-group can be configured individually in the environment of only the backbone atoms using a systematic search procedure combined with extensive local energy minimization. Tests, using the main-chain or both the main-chain and remaining side-chain atoms to calculate low energy geometries for each residue, establish the dominance of the main-chain contribution. The final structure is achieved by combining the individually placed side-chains followed by a full energy refinement of the structure. The prediction accuracy of this approach has been compared to other automated procedures.

Surface Contraction and Expansion Waves Correlated with Differentiation in Axolotl Embryos. III. The Shape of the Fate Map

Natalie K. Björklund† and Richard Gordon*
Departments of Botany*, Chemistry†, Microbiology†, and Radiology*,
University of Manitoba, Winnipeg R3T 2N2, Canada

Manuscript started June 17, 1992.
To be presented at: 1993 Open Problems in Computational Molecular Biology: Third International Workshop, Telluride Summer Research Center, Telluride, Colorado, July 12 - 26, 1993.

Audiovisual equipment needed: overhead projector, slide projector, VHS VCR, blackboard.

Extended Abstract

Why do some types of cancer cells occur when a cellular switch is somehow inappropriately stuck? If you cut off a salamander's leg, it simply grows a new leg. How does it do this? Why don't our own cells retain this ability? Why should gene expression in a cell culture sometimes change with generation number? We think that the answer to each of these questions might be found in a complete understanding of normal embryonic development. If we could explain precisely how a single egg can become a multicellular organism we could then use this knowledge to understand many other related cellular processes.

The discipline of embryology has hitherto been to a large degree unsuccessful in achieving this understanding. Embryologists still cannot explain how a cell "knows" that it is to become a certain type of cell at a certain time and in a certain place. We have recently made a discovery that may provide the long missing spatial component of embryological explanation, and turn this situation around.

Imagine yourself as a cell in a urodele (salamander) blastula just prior to the midblastula transition. At this point you are one cell in a hollow ball of cells that up until now has been functioning exclusively using the mRNA left for you from maternal sources. If you are on the bottom of the ball (gravitationally speaking) you have more yolk. If you are on the top you have more pigment. But other than that there is very little difference between you and all your sister

cells in the hollow ball. If you are moved from one end of the ball to the other you won't "mind", you'll just carry on in that spot participating as if you had always been there. (This makes you part of a "regulating" embryo.)

Gradually over the next few hours you will begin synthesizing your own set of mRNA. You will begin to somehow "sense" where you are relative to the other cells in the hollow ball. By the time gastrulation begins, you will have already figured out your relative position enough to have become one of three basic embryonic cell types: endoderm, mesoderm, or ectoderm. If you have become endoderm or mesoderm you will move into the interior of the embryo during the next developmental stage, gastrulation. By the end of gastrulation you will have become either part of a strictly nutritive endoderm, or the endoderm portion that will form the innermost organs. If you are mesoderm you will have determined that you should become either somitic, lateral, notochordal or germ cell line mesoderm. If you are ectoderm, you will have "made the choice" of whether to become neural plate or epidermis. Whatever your position, you will be producing the appropriate mRNA for cells in that position.

If you are in the ectoderm you will stay on the outside for now. If you are on the top hemisphere you will participate in the formation of the future nervous system. If you find yourself the bottom surface, you will form epidermis. If you are roughly in the middle you will become part of the special sense plate eventually forming the sense organs. Of course, if you were neural ectoderm during gastrulation but subsequently found yourself on the interior of the embryo due to neurulation movements, you will somehow know you should change your ways completely and begin forming all the cell types required in the tail. You'll do that even though it may mean (seemingly) reverting back to mesoderm and starting over again to make the inside of the tail.

As a cell you must have some way of keeping track of what your location is now and what it has been. Somehow this three dimensional data must be translated into a form, presumably using repressors, enhancers, methylation, etc., that your one dimensional linear DNA can both record and respond to. Your data storage and retrieval system is so consistent that even while you were a cell in the blastula stage, an embryologist can know what you will become simply by checking your location on a "fate map". Your original location on the hollow ball has been marked with a vital dye and tracked. In this way, all the important tissue types of the embryo can be shown to have originated in the surface cells of the blastula. How you know where you are, what the "developmental time" is, and what

you do next, is a secret you've been keeping from embryologists for about 150 years now. How do you do it? Why do you follow a peculiar path over the surface of the embryo described by the strange shape of the fate map?

Now let's go back to the embryologist studying this problem. How much can be explained? There are quite a few things known for certain. Throughout the earliest stages of development, cells experience a special time when they are "competent" to sense their location in the embryo. This "competence" is transitory. If the message, the "induction", comes before they are competent, the cells can't respond. But while they are competent they can be induced by everything from boiled Guinea pig's liver to activin. Similarly, once a cell is set on a particular pathway to a certain final cell type, it is usually impossible to change it to follow a new path. Obviously understanding the nature of "competence" and the signalling system of "inductions" cells respond to, would explain a lot about how cells do what they do.

But a specific chemical inducer has never been found in spite of three generations of long and intense investigations. Rather, a bewildering assortment of both artificial and natural inducers of ectoderm into neural tissue is known to exist. In fact, the eminent embryologist Dr. Pieter Nieuwkoop has been heard to state:

"Anything you can find in your kitchen garbage will probably induce ectoderm."

In despair, the embryological community largely gave up on explaining primary neural induction of ectoderm and turned their efforts to other tissues (especially mesoderm) instead. Sadly for the chemical inducer proponents, mesoderm also has a rapidly growing list of possible inducers.

Beth Burnside (1971, 1973), using transmission electron microscopy, found that in the apical end of each ectoderm cell there is a cytoskeletal apparatus. It consists of a microfilament ring with microtubules lying in a mat coplanar to the ring. In our own laboratory we have confirmed her result and also found that immediately below the microfilament ring is an intermediate filament ring.

We have also discovered a number of surface contraction and expansion waves that appear to correlate with each of the steps of differentiation in the amphibian embryo, (Gordon, Björklund, & Nieuwkoop, 1994). The waves are known to occur in the urodele embryo the axolotl, Ambystoma mexicanum, Gordon & Brodland (1989) and the fruit fly, Drosophila. (Poodry, Hall & Suzuki, 1973; Suzuki, 1974;

Ready, Hanson & Benzer, 1976). The waves themselves propagate from one cell to the next as a visible contraction or expansion of the apical surface of each cell. The cells involved in a contraction wave remain contracted for about ten minutes and then they relax. The relaxation probably occurs because the intermediate filament ring provides an elastic component to the cell (Brodland & Gordon, 1990). The cells involved in an expansion wave appear to simply expand and remain expanded.

The waves are paired. A single early tissue type will eventually differentiate into two daughter types each with it's own wave. For example, the portion of the ectoderm that experiences a contraction wave becomes neural tissue. The portion that experiences an expansion wave becomes epidermis. We now know that each of the final tissue types that is known to exist by the end of gastrulation experiences a unique sequence of expansion and contraction waves. By tracking the fate map as it invaginates, we can show how the fate map's peculiar shape is directly correlated to the sequence of contraction and expansion waves we have discovered.

In Gordon & Brodland (1987) it was postulated that the microfilament ring and the microtubule mat in the apical end of each competent ectoderm cell is in a state of balanced mechanical equilibrium. When the pharyngeal endoderm touches the ectoderm, the force on these ectoderm cells due to the internal pressure of the hollow ball is reduced. The reduction in the outward force gives a mechanical advantage to the microfilament ring. The cell contracts. It tugs on the adjacent cell and, like the stretch activated contraction of smooth muscle, the adjacent cell contracts in response. The mechanical signal to contract is then passed through the upper portion of the ectoderm as a self-propagating signal. On the other side the microtubules get a mechanical advantage. This sets off a wave of expansion. We do not know yet why the waves propagate over one region of a tissue type. We suspect that simple mechanically based restrictions will explain this.

For our proposed signalling system to work, cell state splitter construction is matched by the preparation of two possible signals for triggering one of two gene cascades. Once the cell contracts or expands a signal indicating the type of wave is sent to the nucleus. The nucleus responds to the signal by initiating one of the two possible gene cascades. The end of the gene cascade includes the resetting of the mechanical instability (competence) by the construction of a new cell state splitter, two new signalling systems, and two new gene cascades. The cell waits for the next wave to come along, participates in it, and propagates it.

In this model, a cell does not "know" where it is. What it records (in some

manner yet to be determined) is the sequence of contraction and expansion waves it has experienced. One prediction is that all cells that have experienced the same sequence of waves are developmentally equivalent. This contrasts with the concept of "positional information", in which each cell knows where it is by some mechanism or other, and uses a "lookup table" to determine what to do next.

The nature of the signal between the contracting or expanding cell state splitter and the subsequent gene cascade has been postulated in Björklund & Gordon (1993). Because of the many studies done on the nature of competence and induction there is a huge body of literature on what happens to cells that are induced to become neural tissue from ectoderm. For example, it is known that these cells experience a protein kinase C translocation followed by a rise in cAMP levels. The PKC translocation is prolonged, lasting about ten minutes. Further, different isozymes of PKC are known to occur at different stages of development in different tissues. On the basis of this we suggest that there is a calcium wave propagated from cell to cell causing both the microfilament ring contraction and the PKC translocation. We envisage two master genes prepared to respond to a signal generated from the PKC translocation. One master gene is triggered and it sets off a gene cascade appropriate to neural tissue. The other master gene would require a different signal to be triggered. Excellent candidates for generating the signal are the microtubule associated proteins. If the microtubules are suddenly greatly expanded in an expansion wave the number of MAPs that bind to these microtubules would also be suddenly increased. Their activity could then trigger a gene cascade appropriate to epidermal tissue.

In our presentation we will show our computer generated time lapse images of contraction and expansion waves. We will show the results of our laboratory's EM work on the cell state splitter. We will explain how the peculiar shape of the urodele fate map can be explained by the succession of waves. We will present a differentiation tree showing the sequence of expansions and contractions that each tissue type present by the end of gastrulation has experienced. We will introduce our ideas for the signalling system between the physical waves we see and the gene cascades the nucleus responds with. Finally we hope to discuss the implications of the waves to other specialties including molecular biology, embryology, and evolution.

References

Björklund, N. K. & R. Gordon (1993). Nuclear state splitting: a working model for the mechanochemical coupling of differentiation waves to master genes.

Ontogenez (2).

Brodland, G. W. & R. Gordon (1990). Intermediate filaments may prevent buckling of compressively-loaded microtubules. J. Biomech. 112(3), 319-321.

Burnside, M. B. (1971). Microtubules and microfilaments in newt neurulation. Dev. Biol. 26, 416-441.

Burnside, M. B. (1973). Microtubules and microfilaments in amphibian neurulation. Amer. Zool. 13, 989-1006.

Gordon, R., N. K. Björklund & P. D. Nieuwkoop (1994). Dialogue on embryonic induction and differentiation waves. Int. Rev. Cytol. in press.

Gordon, R. & G. W. Brodland (1987). The cytoskeletal mechanics of brain morphogenesis: cell state splitters cause primary neural induction. Cell Biophysics 11, 177-238.

Poodry, C. A., L. Hall & D. T. Suzuki (1973). Developmental properties of *shibire* ^{ts} : a pleiotropic mutation affecting larval and adult locomotion and development. Dev. Biol. 32, 373-386.

Ready, D. F., T. E. Hanson & S. Benzer (1976). Development of the Drosophila retina, a neurocrystalline lattice. Dev. Biol. 53, 217-240.

Suzuki, D. T. (1974). Behavior in *Drosophila melanogaster*: a geneticist's view. Can. J. Genet. Cytol. 16, 713-735.

Deriving Non-Homogeneous Markov Chain Models from the Multiple Alignment with the Entropy Criteria.

Mark Borodovsky,
School of Biology
Georgia Institute of Technology
Atlanta, GA 30332-0230

Many biologically significant regions of the DNA (and protein) sequences do not reveal any consensus-like pattern which would be determined by the traditional multiple alignment algorithm based on matching scores. The important example of that kind would be the DNA protein-coding regions for which the attempt of the multiple alignment might be successful only in the case of collecting a narrow group of closely related genes.

The general idea to perform multiple alignment for several randomly chosen gene sequences looks unreasonable. Nonetheless, when it is placed into a bit different context this idea leads to the extracting a non-homogeneous Markov chain model of the protein-coding sequence. This model might be considered as a generalization of the consensus or profile-like model of the DNA functional region.

We suggest a multiple alignment method for deriving the parameters of the non-homogeneous Markov chain model from the sample of sequences which are suspected to share common compositional pattern and which can be described by virtue of that model. Usually such a pattern might be expected if there is an a priori known phasing of the biologically active DNA (like triplet genetic code), positioning around the firmly established single point (like origin of replication or transcription) and soon.

The main step of the method's algorithm is the Shannon entropy calculation procedure which is applied to each one column in the multiple alignment "box" which width is chosen usually bigger than the length of the expected pattern. The column entropy value is a function of the frequencies of oligonucleotides associated with the given column (gaps are not allowed). The entropy alignment criteria (for the whole "box") is obtained by summing the column entropy values up. This entropy total is storing, sequences are shifting according to a certain (random or deterministic) rule and the new calculation of the

entropy criteria follows until the minimum value is achieved. The final configuration of the multiple alignment is the one which is used for the definition of the non-homogeneous Markov chain model parameters (transition probabilities).

The results obtained currently for the set of sequences generated by various types of Markov chain generators show that the method allows to align sequences in the "box" and extract the hidden phase from the set of DNA fragments which were randomly shifted in the beginning. In the case of real protein-coding sequences the method gives the result which clearly indicates the triplet phase and allows to extract the non-homogeneous Markov chain model which has already been used for the purpose of gene prediction in the GENMARK algorithm (1-3). Some other sets of nucleotide and protein sequences are now under consideration.

There is one remarkable property of the entropy minimization based multiple alignment. The final alignment configuration leads to such a Markov chain model which defines the maximum value of the likelihood that sequences (in the box) would appear as an output of the generator defined by this model. This maximum value is determined in comparison with the values obtained from the similar model extracting from any
other intermediate alignment configuration.

1. Borodovsky M., and McIninch J. (1993) Computers & Chemistry, to appear.

2. Plunkett G. III, Burland, V., Daniels D.L., and Blattner F.R. (1993) Nucleic Acids Research, to appear.

3. Noble J.A., Innis M.A., Koonin E.V., Rudd K.E., Bunuett and F., Herskowitz (1993) PNAS, to appear.

THE LIMITS TO REASON
What Science Can Know About Everyday Events

John L. Casti

Technical University of Vienna
Vienna, Austria

AND

Santa Fe Institute
Santa Fe, NM 87501

The primary goal of science is to offer convincing answers to the question: ``Why do we see what we do and not see something else?''. In confronting this question, science is distinguished from its many competitors in the reality-generation game by the particular sorts of methods and tools the scientist employs. The *scientific* answer is a set of rules, i.e., an algorithm, usually encoded as a mathematical model or computer program, with which one can explain the observed phenomena and predict (sometimes) what will happen next.

This talk explores the possible limitations of rule-based procedures for reality generation. In particular, a number of processes from everyday life, ranging from weather and climatic changes to stock market price fluctuations and even to the outbreak of warfare, will be examined in an attempt to determine the degree to which the science of today is in a position to give a convincing set of rules for predicting and/or explaining such phenomena. The lecture concludes with some general ideas centered around the Turing-Church Thesis and the theorems of Goedel and Chaitin for why we can never expect to achieve perfect prediction and explanation---scientific-style---of any natural or human activity.

Estimation of Protein Coding Density in a Corpus of DNA Sequence Data

James W. Fickett and Roderic Guigo

Theoretical Biology and Biophysics Group and
Center for Human Genome Studies
Los Alamos National Laboratory, Los Alamos, NM 87545

ABSTRACT

A number of experimental methods have been reported for estimating the number of genes in a genome, or the closely related coding density of a genome, defined as the fraction of base pairs in codons. Recently, DNA sequence data representative of the genome as a whole have become available for several organisms, making the problem of estimating coding density amenable to sequence analytic methods. Estimates of coding density for a single genome vary widely, so that methods with characterized error bounds have become increasingly desirable. We will present a method to estimate the protein coding density in a corpus of DNA sequence data, in which a "coding statistic" is calculated for a large number of windows of the sequence under study, and the distribution of the statistic is decomposed into two normal distributions, assumed to be the distributions of the coding statistic in the coding and noncoding fractions of the sequence windows. The accuracy of the method is evaluated using known data and application is made to the yeast chromosome III sequence and to C. Elegans cosmid sequences. It can also be applied to fragmentary data, for example a collection of short sequences determined in the course of STS mapping.

INTRODUCTION

Fundamental knowledge about an organism includes an estimate of the number of genes in its genome -- one measure of the overall complexity of the organism -- and an estimate of the closely related coding density (defined as the fraction of base pairs that are in codons). The latter is a basic aspect of genome structure, related to the intriguing question of the prevalence of "junk" or "selfish" DNA [Orgel and Crick, 1980]. An estimation of coding density has important practical consequences as well, for example in deciding whether more information will be gained by sequencing cDNAs or genomic DNA. Current estimates of coding density for most eukaryotic organisms are given

only in rather wide ranges. For example, Clark et al. [1988] estimate that there are roughly 3500 essential genes in Caenorhabditis elegans, giving both reasons to think this estimate may be too high as well as reasons that indicate it may be minimal. Combined with the results of Park and Horvitz [1986], which suggest that half of the genes in C. elegans may be inessential, this gives an estimate of roughly 7000 genes. However Waterston et al. [1992] estimate that the true number of genes in C. elegans may be closer to 15000.

Kaback, Angerer, and Davidson [1979] showed that roughly 50-60% of the yeast genome is transcribed in roughly 5000 transcripts, under laboratory conditions. (A transcript density of 50-60% corresponds to a coding density of less than 50%, since an mRNA contains untranslated regions.) If all genes were distributed uniformly over the genome, this estimate would give about 120 genes on chromosome III. However Yoshikawa and Isono [1990] found 156 transcripts from chromosome III, and Oliver et al. [1992], equating genes with open reading frames of length at least 100 amino acids on the chromosome III sequence, find 182 genes, giving a coding density estimate of 67% and an estimate of about 8000 genes in the whole organism.

The analyses of Waterston et al., Sulston et al. [1992], and Oliver et al. were made possible because an important new source of data has recently become available. Whereas most sequences in the current databases are from highly expressed genes, sequence is now becoming available which is a much less biased sample of the genome. In some cases this means a very long stretch of DNA encompassing many genes, as in the case of the recent determination of yeast chromosome III [Oliver et al. 1992], in others it means a large number of short sequences, randomly selected from the genome in the course of determining STSs for genome mapping [Olson et al. 1989].

Current methods to determine coding density, both experimental and computational, rely on counting genes. The experimental methods typically give low estimates because, under the experimental conditions chosen, not all genes are required or expressed. A major difficulty with the computational methods applied to date is that current gene recognition methods have rather large, and sometimes uncharacterized, error rates. Thus Oliver et al. simply count open reading frames exceeding a certain size and the studies of Waterston et al. and Sulston et al. depend on the gene recognition program Genefinder. In neither case is it easy to evaluate the accuracy of the predicted number of genes.

We will present a method to estimate coding density in a corpus of sequence data -- either long stretches of sequence data or a large number of short sequences -- that does not rely on identifying genes. Indeed, it is possible to define on windows of sequence a number of simple measures, or coding statistics, which are indicative of protein coding function (reviewed in [Fickett and Tung 1992]). And while such statistics have a large random component and a large variance when observed on individual windows, the overall distribution of such a statistic, when observed on a large set of windows, is closely correlated with global coding density (Fig. 1).

One simple approach to the problem considered here would be (1) to infer, using sequence data of known coding density from public databases, a model of the relationship between coding density and an ensemble property of the coding statistic distribution, as for example the linear regression shown in Fig. 1, and (2) use such a model to predict the coding density of the new sequence data under study. However, because the data in the public sequence databases is a very biased sample of the genome, extrapolation from the database to the genome may not be justifiable.

This problem may be surmountable, but here we pursue an alternative method which does not rely on first establishing a model of the relationship between coding density and a coding statistic in previously characterized sequences, but rather depends exclusively on the distribution of the coding statistic on the corpus of sequence data under study. In the method we will present, the sequence data under study are first partitioned into a set of fixed-size windows, and the chosen coding statistic is calculated on each. Then the distribution of the statistic is decomposed into two normal distributions that are assumed to correspond to the distribution of the statistic in the noncoding and coding fractions of the sequence data.

Figure 1. Correlation between the mean of a coding statistic and the coding density calculated from GenBank annotation. For each density d, d=0,..., 100, a set of yeast 240 bp sequence windows was made by selecting windows at random from a GenBank reference set with probabilities chosen to obtain an expected overall percentage d of coding windows, and an expected set size of 315 kb. Each such set is represented as a point in the figure, the abscissa giving the average of the coding statistic (maximum of a codon usage discriminant function over the six frames of the window) over the set, and the ordinate giving the coding density computed from GenBank annotation

We will first introduce the Max Codon Usage coding statistic, or MCU, for which we have generally observed reasonably gaussian behavior. We will describe in detail the method used to decompose the distribution of the coding statistic, using the yeast genomic sequences from GenBank to explicitly illustrate it. We will then evaluate the accuracy of the method in large sets of characterized genomic sequences from five distantly related genomic organisms, and describe two applications. In the first, the coding density of yeast chromosome III is estimated by decomposing the distribution of the MCU distribution. In the second, the coding density of a collection of sequenced cosmids from C. Elegans is estimated. In this case, however, since we strongly suspect that the MCU statistic does not have a normal distribution, the estimate will be obtained by decomposing the distribution of a different coding statistic. Finally, we will discuss the applicability of the method to other genomes, and its limitations.

REFERENCES

Clark, D.V., Rogalski, R.M., Donati, L.M., and Baillie, D.L. (1988) Genetics, 119, 345-353.

Fickett, J.W. and Tung, C.-S. (1992) Nucl. Acids Res., 24, 6441-6450.

Kaback, D.B., Angerer, L.M., and Davidson, N. (1979) Nucl. Acids Res., 6, 2499-2517.

Oliver S.G., et al. (1992) Nature, 357, 38-46.

Olson, M., Hood, L., Cantor, C., and Botstein, D. (1989) Science, 245, 1434-1435.

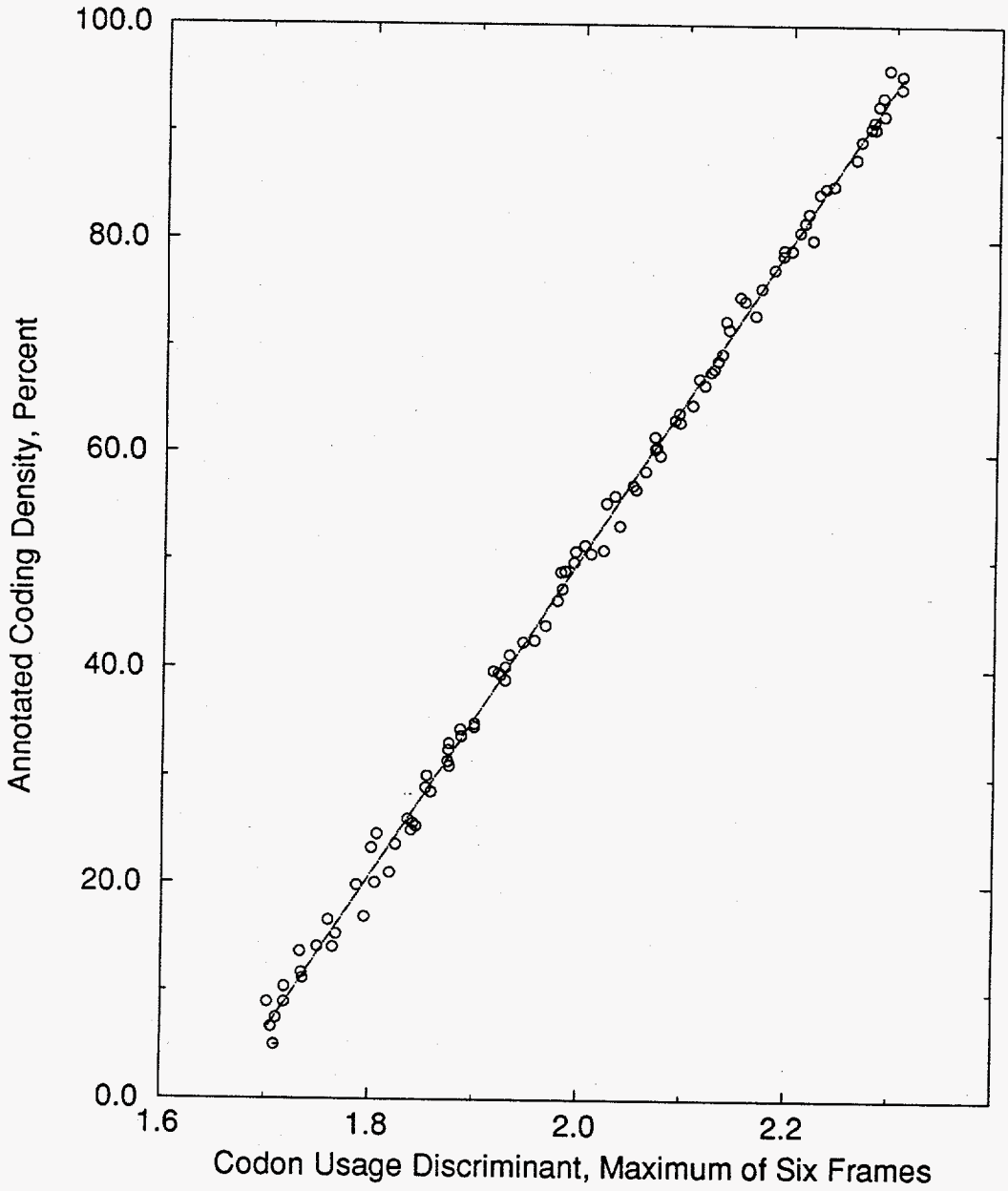
Orgel and Crick (1980) Nature, 284, 604-607.

Park, E.-C. and Horvitz, H.R. (1986) Genetics, 113, 821-852.

Sulston, J., Du, Z., Thomas, K., Wilson, R., Hillier, L., Staden, R., Halloran, N., Green, P., Thierry-Mieg, J., Qiu, L., Dear, S., Coulson, A., Craxton, M., Durbin, R., Berks, M., Metzstein, M., Hawkins, T., Ainscough, R., and Waterston, R. (1992) Nature, 356, 37-41.

Waterston, R., Martin, C., Craxton, M., Huynh, C., Coulson, A., Hillier, L., Durbin, R., Green, P., Shownkeen, R., Halloran, N., Metzstein, M., Hawkins, T., Wilson, R., Berks, M., Du, Z., Thomas, K., Thierry-Mieg, J., and Sulston, J. (1992) Nature Genetics, 1, 114-123.

Yoshikawa, A., and Isono, K. (1990) Yeast, 6, 383-401.



SCIENCE IS IN THE EYE OF THE BEHOLDER

Richard W. Hamming

Introduction

The title is an obvious variant of the standard remark that "Beauty is in the eye of the beholder", and clearly implies that I do not think there is a single scientific method, but rather that there is only some degree of agreement among various practitioners of science. How did I come to this view? That is part of the talk, and by telling you stories about my experiences I hope they will bring you to view the matter somewhat as I do.

There is a second reason for the talk. It was obvious to me from the start that I could tell you little of use to you about your own field of specialization. What I could do is emphasize the famous remark of Pasteur,

"Luck favors the prepared mind."

The more years I have contemplated his remark, and examined the history of science, the more I believe in it. Yes, the particular thing you do is to a fair extent a matter of luck, but if you prepare yourself in many ways then it is reasonable that a great success should strike you. The multiple successes by the same person in the history of science are too many to believe much in the "random luck theory".

As you know beauty in a woman of the Ubangi tribe means very large, distorted lips, around the year 1000 Japanese women blackened their teeth, and currently American women paint their lips a brilliant red and often color their finger nails vivid colors - all in the name of beauty. Thus the standard of beauty, and by inference, of science, depends on the particular social group you are a member of, as well as the particular age in which you live. There is not a single scientific method practiced by all scientists, there is only a vague culture in which you happen to operate. People from different fields tend to cling to the standards of doing science that they were raised on, and by inference try to impose them on others.

By the time half a dozen different teachers in school had told me about the scientific method I noticed that they did not say the same things, rather they differed a good deal. They all emphasized the importance of first getting

measurements, data if you prefer, and then making the theories. As Sherlock Holmes said, "One should not theorize before your data." By the late high school days, and more clearly in college, I realized that you had to have some sort of theory to tell you what kinds of data to gather. But also that you had to have some data before you could have any theory! Neither theory nor practice could completely precede the other. It could not be a simple process! Theory and experimental data must go hand in hand.

From a lifetime of watching, and asking people about what they were doing, I have come to the realization that in my areas of expertise, mainly the so-called hard sciences, there is a gradual process of having some sort of hunch, gathering some data which may confirm it, making a more careful, and possibly somewhat altered theory, followed by more data gathering, both more extensive and often more careful measurements in the lab. This in turn generates more theory and more data gathering, perhaps widening or narrowing of the area of application, the range of use, etc., until one has a decent theory of some generality.

Any theory you create should be followed by an active, careful search for why your final theory might be wrong. This last stage is, in my opinion, an essential part of doing science. Unfortunately in this day and age of rapid publication this luxury is rarely indulged in - whoever publishes first, if they are right, gets the credit and the careful, cautious person is left out completely! And if they are wrong, in many areas this is soon forgotten, if ever noticed at all! The idea of doing careful science is rapidly fading under the pressure of publish or perish, to the detriment of science generally.

Gather the Data First

In the softer sciences people tend to start with volumes of data from which they think they will extract a theory. To illustrate the extreme, suppose I make random entries for 100 measurement of 20 variables. I now, as they do in some areas, let the computer go to work and compute every one of the 190 correlation coefficients. With a high probability I will find one or more significant correlations. But since the data was random one believes that there is nothing there at all! That is what is wrong with this simple model of science which says that you gather the data, search for a theory, and then publish. Such science is apt to be ephemeral!

In a conversation with Einstein after a talk he had given, Heisenberg observed that he had followed the method Einstein had used when he had created the special theory of relativity. Einstein laughed and replied, "I may have used it but still it is nonsense." He went on to explain, "Whether you can observe a thing or not depends on the theory which you use. It is the theory which decides what can be observed."

Let me illustrate this with an example you all know. Newton had suggested that light might consist of particles (he knew of course about Newton's rings), and this became, ignoring Newton's caution, the dominant theory of light until the wave theory arose. In this new theory the raw facts were seen differently. The edge no longer attracted the particles and caused the diffraction pattern. Where once the particle on hitting the photographic plate developed the grain of silver, now it was not clear how the wave got all its energy to one place to apply it to the single grain of silver salt. Of course at the moment we have in Quantum Mechanics a lovely combination of the two theories, but they do not mix very well at all. As your professor of QM had to say with regard to the two slit experiments (which Feynman insists is at the bottom of QM), "I can't explain it, you will get used to it."

There are more arcane theory changes in which literally the facts changed with the change of theory, but this is not the place to go into them. I suspect that is you search your own knowledge of your field you will find examples of where when you got a new theory the older facts seemed to change radically. There is a lot of truth in Einstein's remark that it is the theory that determines the facts.

Ockham's Razor

There are many criteria in science. For example Ockham's razor which says, in modern words, that you do not make unnecessary assumptions. You keep your assumptions down to as few as possible. It sounds like a good rule, and it ought to keep the danger of contradictory assumptions to the minimum.

In a graduate course in abstract algebra I noticed that most text books gave three conditions for a group, but there was one book which gave only two. The two conditions required a lot of hard preliminary proofs to show that they implied the three assumptions usually given. Furthermore, the two assumptions were so peculiar as to be hard to understand while the three were easily understood. From that, (and there were a couple of other similar cases in

abstract algebra), I came to the belief that it was the clarity in understanding the assumptions that took precedence over Ockham's razor. But, as we say in the military design of weapons field, I know of no exchange ratio to tell me how much to give up of one for a gain in the other, nor can I measure ease of understanding in an objective way.

Accuracy of the Theory

We often hear that the theory which gives a better fit to the existing data is the better theory. But we all know that the theory Copernicus proposed did not fit the data, as he used it, as well as the older Ptolemy predictions, yet in time the Copernican theory displaced the theory of Ptolemy and is often regarded as one of the significant steps forward in science.

It should be obvious that at the start of any new theory it cannot account for as much as the older theory; for a long time Quantum Mechanics could not explain as much as the classical theory, yet it has won out!

In the early days of a new theory we usually try applying it to all kinds of situations, and thus we gradually learn the limitations of the theory. A lack of accuracy in the early stages of a theory may mean only that we have not learned the limitations of the theory, or it may mean that we have to modify things a bit - which to do is not obvious at the time. And, of course, it may mean that the theory should be abandoned, or maybe the data is bad!

Controlled Experiments

Another criterion of science one often hears is that you must be able to do controlled experiments to be sure of what you are observing. Again this is plausible, but Astronomy for most of its existence had absolutely no control of what it was observing; all that astronomers could do was to choose the time and place of observation. There was no possibility of repeating exactly the same experiment, or repeating with all but one factor the same. But most of us will agree that Astronomy was the first of the sciences in spite of this handicap. Hence controlled experiments cannot be an essential feature of science.

In the current cosmological theories there is precious little checking with facts, and in some of the theories the essential parts may never be testable!

A Theory Must be Potentially Disprovable

Karl Popper has made this a popular criterion. He came to it apparently because while he was studying psychoanalysis in Vienna he gave a psychoanalyst a set of symptoms and got an explanation, another set of symptoms and another explanation - and suddenly, like the proverbial light bulb in the comic strips, he realized that a theory that could explain any set of data could explain nothing! A theory which can explain everything can explain nothing. If there is no conceivable set of data which would vitiate a theory then it is not science, so Popper claims. Again, it is not bad criterion to keep in your mind - what kind of data could prove you to be wrong - but it is not an absolute standard to meet at all times.

Consider the history of gravity waves. The attempts to measure them has now a long history of significantly more and more accurate measuring instruments (of higher and higher cost) and has produced nothing. A few years ago I heard of an experiment being done in Italy which is supposed to have 100 times or more accuracy than any previous attempt, but I think they had better begin to think about what they will do if once more they do not find anything significant. What would they accept that there are no gravity waves? That their theory could be wrong? Popper is looking over their shoulders! When you keep looking for some effect and it keeps getting smaller and smaller and harder and harder to detect, when do you decide that there is nothing there? There is a classic paper on the dangers of working at the edge of detectability.

A Theory Should not Involve Unmeasurable Phenomena

As you all know we once had a theory of an ether that filled all space and was the medium through which light waves propagated. But it proved to be unmeasurable by anything we could imagine, and we gradually, for most people, abandoned it.

To what extent the ether is now back among us in a similar form through which is propagated the low temperature background radiation left over from the supposed "Big Bang" is difficult to say at the moment.

Maxwell, when he wrote his two famous papers on the kinetic theory of gasses, had to assume molecules for which there was no possibility then of seeing, measuring directly, or doing any other thing, to make them real; they

were simply fictitious entities to explain other observations. Yet, the success of his theory brought us to believe in molecules, though the hard boiled people who believed only in measurable things resisted into the 20th century! This attitude of considering only observables (through, of course, suitable instruments), was popular in the Vienna circle of positivists, and in fact Einstein in his early years believed in it, as did Heisenberg, but in time both seemed to have abandoned this strong position of dealing only with measurements and eschewing imaginary entities (see above story).

Fruitfulness

A valuable property of a theory is how much it suggests things to do that you had not thought of earlier. A complete, closed theory is not much use; it will not lead you to the future theory that is going to displace, in time, the current one you are working with. I once asked Walter Brattain, of the transistor Nobel Prize, why they had succeeded when others who working at the same time on the same general idea failed. His simple answer was that they succeeded because they had kept theory and practice (experiments) together, and when one side was stuck the other suggested things to do. It is not so important to have the correct theory, or even exactly the correct measurements, as it is to have a basis for further action when you get stuck.

As a partial example of the application of this consider the following story. When Bell Tel Labs was finally successfully making the early point transistors, and Shockley was starting on junction transistors, a V. P. is reported to have said, "I will back the junction transistor over the point transistor because the scientists think they can understand the junction transistor." He clearly expressed his faith in the relative fruitfulness of the theory for the junction transistor.

Top Down vs. Bottom Up

One scientist I recently talked to about his style of doing science claimed that he analyzed things carefully and tried to find the central problem, and fitted other things around it. I observed that some scientists do exactly the opposite, they start almost anywhere and mess around until they remove all the difficulties and then have the result they want. In writing computer programs the two extreme approaches are called "top down" and "bottom up"; in practice most people use a combination of both methods - neither extreme is as good as

a judicious mixture applied properly, and what you do may depend on the particular situation you face at that time.

Unifying Power

A characteristic of important theories is that they unify what before were seen as separate things - that they have breath of application rather than a narrow, sort of ad hoc, structure. Consider the great unification that Maxwell's equations made! The breadth of application gives some "truth" from each part.

Mathematics is, of course, the great example of this unification of different phenomena via the same equations. I will discuss mathematics later.

Save the Phenomenon

There is a wide spread belief, especially among the philosophers of science (who generally have never done real science), that all science is merely a mnemonic for connecting various results in a framework so that they can be retrieved easily. They maintain that there is absolutely no "truth" in any theory. On their side is the obvious fact that all previous theories have been rejected to make place for the present theories, and we can hardly suppose that we have now reached the millennium of unchanging theories.

Against them is the simple fact that most scientists talk and act as if they believed their current theories represented "reality", perhaps not "exact reality" but close to it. By believing that our theories represented "reality" we have made much progress, and successfully predicted new, unknown effects. If the theories were not representations of "reality" how does this happen so often? But, again, we know in our hearts that our theories are doomed to be replaced. The phenomena to be explained stays much the same but the theory may change greatly.

For example in the Middle Ages the belief seems to have been that the angels pushed the planets around the heavens. They seemed not to have recognized inertia clearly so something had to keep the planets moving, (note that it must have been a deadly monotonous job for the angels). Later Newton gave formulas describing how inertia kept them moving without any other forces, and how gravity bent the trajectories into ellipses. In the theory of General Relativity they claim that the mass "bends the space" to produce the

observed trajectories. Three radically different theories, but pretty much the same phenomena were explained by them.

There is not a unique theory

There is not a unique theory to account for, or "explain" a set of data. In QM we started with the Heisenberg Matrix Mechanics and the Schrödinger Wave Mechanics, which, along with the later Group Theory approach, were shown to be, in a limited sense, equivalent.

While supervising a Ph.D. thesis I acquired from another professor I found that in his area of research, using mainly random inputs with selected power spectra and measuring the outputs of the black box one could determine an internal structure, but that this structure was in no way unique! No possible set of measurements of the kind being made, could ever distinguish between two radically different insides of the black box. Similarly, from a set of data you cannot hope to prove a unique theory is correct. As in QM there can be multiple theories which agree on the measurements, but have different theoretical foundations.

And you all know that while you may say carelessly that Euclidean and analytic geometry are the same, in practice there is not a large amount of overlap - we use each where it is most convenient, and would hesitate to prove some results obtained by one method by the other method. For example, Euclidean geometry proves the Pythagorean theorem, analytic geometry assumes it!

Mathematics

There is a widespread belief that the more mathematics that there is in a field the more "scientific" it is. This belief is usually based on the idea that mathematics is absolutely certain knowledge. This is certainly what the early Greeks believed. But Kline has written a book Mathematics: The Lost of Certainty in which his aim is to show how we have passed from the Greek belief to our present one that there is no "truth" in all of mathematics.

There are five different schools of the philosophy of mathematics: (1) Platonism, where ideas are eternal and are the only truth; (2) Formalism, where we abandon all meaning and merely manipulate symbols according to

arbitrary rules, (when rigor enters meaning departs), (3) the Logical school which attempts to show that all of mathematics is a branch of logic; (4) the intuitionists who believe that in a field where we have a rising standard of rigor there can be no final proofs, and at the bottom is human intuition not logic; and (5) the constructivists who insist that things must be constructed to show that they "exist". But the constructivists take away so much of mathematics as we use it that only a few computer people tend to like the school. None of the schools has succeeded in winning the majority of people who think about what mathematics is.

You have often heard things like: nothing is more sure than that 1 + 1 = 2, yet I wrote a book in which the bulk of the arithmetic and algebra had the rule 1 + 1 = 0. Of course you will say that the 1 in the two equations is not the same thing, just as the points, lines and planes in Euclidean geometry and non-euclidean geometry are not the same things.

At the end of the Middle Ages when people began to face the topic of forces they found that the conventional, received from the past, arithmetic did not work when adding forces, and they had to create a new kind of mathematics called vectors. As just noted, to develop error correcting codes I had to use arithmetic modulo 2 and abandon both standard arithmetic and the Pythagorean distance.

Much of current mathematics is tainted with "the whole is the sum of the parts" but it appears to me that in your field often the whole is more than the sum of the parts, that the modern word "synergism" applies. This suggests to me that you will in your turn have to invent the mathematics you need. Sometimes, as was the case with Heisenberg, the mathematics of matrices had already been developed, but I doubt that you will be so fortunate much of the time. But I tell you from personal experience, once you are clear in your mind as to what you are dealing with, then it is not hard to create the corresponding mathematics. After all, according to an ex-department head of mathematics at Bell Tel Labs,

Mathematics is nothing but clear thinking

Hence if you are going to think clearly you are going to be doing mathematics.

When I was considering accepting the invitation to give a talk, it was this

point that caused me to accept. I hoped that I could get you to see that the mathematics you were taught in school is not sacred, that it was not on the stone tablets that Moses brought down from Mount Sinai, but that it is human made to meet conditions that we face. The conditions in your field seem to me to be sufficiently different from those of the past as to require you to create new forms of mathematics.

COORDINATED CHANGES IN PROTEIN MULTIPLE SEQUENCE ALIGNMENTS

K.Hatrick and W.R.Taylor
Laboratory of Mathematical Biology,
National Institute for Medical Research,
The Ridgeway, Mill Hill,
London NW7 1AA,
U.K.
Phone:081 959 3666 ext2396

(1) Introduction

In protein multiple sequence alignments, positions conserved with regard to amino acid type or property are often spotted by eye. In contrast, relationships between pairs of positions (columns) in multiple sequence alignments are best detected by other methods: the term covariance defines one such relationship, where variation as to amino-acid type in one column is echoed by complementary change in amino acid type in the second column. For instance, two columns in a protein family multiple alignment may correspond to residues necessary for salt bridge formation: in such a case, whenever the residue type in one column is positively charged the corresponding residue in the second column will be negatively charged, and vice versa.
If the two columns vary in such a manner for all proteins in the multiple alignment and the pattern of covariation is sufficiently complex, is this an indication of proximity of the two residues in the tertiary structure of a member of the protein family? Previous attempts *(1 ,2)* to answer this question have used exact pattern matching methods and further investigation using site-directed mutagenesis. These attempts have encoded columns in a multiple alignment (IIIVVL) as numbers (111223) by assigning sequential numbers to each amino-acid type. Columns which are assigned the same number (SSSDDM = 111223) are considered correlated. However, when a protein multiple sequence alignment is made up of two or more subfamilies, this method lacks the ability to discern between pairs of conserved columns which show variation between subfamilies and covarying pairs of columns. Conserved pairs of columns with subfamily-specific differences can often be found in the active site or core of the protein, and are quite distinct from covarying positions. In addition, any method to detect covariance between columns must be able to distinguish complementarity between non-varying columns, which does not suggest association of the columns, from covariance between columns which do vary: when one column is conserved with respect to positive charge and another is conserved with respect to negative charge, the two columns are not considered to exhibit covariance, yet both are assigned equivalent patterns (namely 111.....1) by the most simple pattern matching schemes. In this work, new methods for detecting covariance which surmount these problems are presented.

(2)Methods for detection of covariance in protein sequence alignments.

(2.1) Simple column scoring scheme.

Consider the multiple alignment shown in fig.1. Each column in the alignment is a mosaic of black, white and grey squares. Dashes represent indels. Each square represents a residue in the alignment; black squares represent residues with one characteristic, white squares represent residues with the complementary characteristic. Grey squares represent residues which posess neither characteristic. A simple scoring scheme was devised to score one column (i) against another(j):

Here,

a = No. blacks in column(i) matched with whites in column(j);
b = No. whites in column(i) matched with blacks in column(j);
c = No. blacks in column(i) matched with blacks in column(j);
d = No.whites in column(i) matched with whites in column(j).

Then

$$\text{score}(i,j)=\min((a-c),(b-d)) \quad \text{and} \quad \text{score}(j,i)=\min((b-c),(a-d)) \qquad (1)$$

Note score(i,j) does not necessarily equal score(j,i).
The minimization ensures that columns with approximately equal numbers of blacks and whites score most highly if covariance is found.

As we wish to establish a scoring system in which the score of two columns is identical whether score(j,i) or score(i,j) is taken, we define the function

$$\text{total\_score}(i,j) = \text{score}(i,j) + \text{score}(j,i) \qquad (2)$$

Thus the best score for any particular column(i) in an alignment of length N is given by

$$\text{Best}(i) = \max_{j=1}^{N} \text{total\_score}(i,j) \qquad (3)$$

When column(j) gives the best score for column(i) and column(i) gives the best score for column(j), the relationship between column(i) and column(j) is called reflexive. This is only necessarily the case for the best score out of all columns. This is given by

$$\text{Max}_{i=1}^{N} \; \text{Max}_{j=1}^{N} \; \text{total\_score}(i,j) \qquad (4)$$

A C program which obtains a list of the best X scoring column pairs (reflexive or both reflexive and non-reflexive) given a MULTAL(3) alignment is freely available on request from the authors.

(2.2) An amino acid network scoring scheme.

Amino acid relatedness has been represented as a metric distance matrix[4]; a regularized model of these relationships has been captured in a three-dimensional network of amino-acids.The scoring scheme presented above (2.1) considers amino acid residues as members of a particular amino-acid set (specified in [5]). Inevitably, some information is lost in encoding a multiple alignment in terms of sets of physico-chemical properties. Taylor[6] has detailed a method for finding covarying columns which incorporates the network information, avoiding this loss. In this method, all covariance involving residues *a* and *b* (at positions I in sequences m and n,respectively) and residues *c* and *d* (at position j in sequences m and n) is identified as follows:

Given a set of vectors {v} between amino acids, the degree of compensatory change between the amino acid pairs (a,d) and (c,b) can be measured directly from either pair of opposing edges, such that

$$t(ijmn) = |ab - cd| \qquad (5)$$

Thus changes in opposite directions on the "edges" (see fig.2) dominate. An alternate formulation gives more explicit weighting to the direction of the vertical edge vectors through use of their dot product:

$$^{v}u(ijmn) = |ab||cd| - (ab.cd) \qquad (6)$$

Alternately, the horizontal edges can be used to give a score:

$$^{h}u(ijmn) = |ac||bd| - (ac.bd). \qquad (7)$$

Both horizontal and vertical edge scores are used to give an overall score $u(ijmn)$, defined as

$$u(ijmn) = {}^{v}u(ijmn) + {}^{h}u(ijmn). \qquad (8)$$

For a given pair of columns in a multiple alignment, $u(ijmn)$ is calculated for all residues (a, b, c, d). In order to calculate the maximum sum of scores, the sequences are divided into two groups by cluster analysis: where the two groups are denoted A = {a1,.....ap} and B = {b1,.....bq}, with C = {c1,...ck} representing unassigned sequences, the preference P of a sequence cj in C for groups A (^{A}P) and B (^{B}P)is given by

$$^{A}P(j) = \frac{\min_{i=1}^{p} s(cj,ai) + 1}{\max_{i=1}^{q} s(cj,bi) + 1} \qquad ^{B}P(j) = \frac{\min_{i=1}^{q} s(cj,bi) + 1}{\min_{i=1}^{p} s(cj,ai)} \qquad (9)$$

(2.2)An amino acid network scoring scheme.

The function s (in (9)) is the score $u(ijmn)$. The maximum of ^{A}P and ^{B}P (the higher group preference) identifies the group to which cj is joined. Other scoring schemes [6] based on diagonally identical and diagonally similar residues in a pair of columns have also been used in place of s.

(3)Results

Results obtained so far have used method(2.2) to analyse covariance.in multiple alignments. Alignments of protein families were obtained using the program MULTAL [3]. Two controls were used to assess the statistical significance of the mean separation for pairs chosen by the compensation measure $u(ijmn)$: the first control is to compare the mean separation for pairs chosen by the compensation measure against the mean distance over all pairs. The second control is to compare against an equivalent mean separation of pairs selected by the estimated *packing-preference* ,p (i,j). This is given by

$$p(i,j) = ci + cj, \qquad (10)$$

where c is a measure of conservation given the amino acid relatedness table (M) by taking a mean over all pairs of aligned residues in a family of N sequences:

$$cj = \frac{2}{N2 - N} \sum_{j=1}^{N-1} \sum_{k=j+1}^{N} MRij,Rik \qquad (11)$$

For the first control, if $u(ijmn)$ selects column pairs at random, then, over a large number of trials, the mean pairwise distances of the samples should be evenly distributed about the mean distance of all pairs: the overall mean can be taken as a class divider, with the expected number of sample means falling above or below this value following the binomial distribution. Where N column pairs are selected by the network scoring scheme, the chance of M or more being selected is

$$PM = 1 - \sum_{M}^{N} 2^{-n} \qquad (12)$$

For the 16 proteins that constitute the data set, any measure that selects 14 or more is unlikely to be a random process. There is a 1/10 chance that 11 will be selected, a 1/26 chance that 12 will be selected and a 1/100 chance that 13 will be selected.

The results for a test set of 16 multiple alignments (each of which contained one protein sequence with a known three-dimensional structure) are shown in table 1.

(4)Discussion and conclusions

Two methods have been presented for the analysis of covariance in protein multiple sequence alignments: both methods avoid the shortcomings of the previous pattern matching method developed to analyse covariance. Results are shown for the second of these methods, expected to be more discerning as regards coordinated change in amino-acid columns.
Table 1 shows that while the mean beta-carbon distance for covarying columns is almost always smaller than expected, conserved pairs (which do not covary) provide a better estimate of amino-acid proximity than the amino acid network scoring scheme, being significantly closer to one another than covarying amino acid positions.Important explanations for the marginal significance of the samples selected by the amino acid network scoring scheme include (a) long range interactions between covarying column pairs, (b) residual bias towards better conserved pairs among the compensated pairs. These factors appear to be less problematic for the automated search for Watson-Crick covariation used for the detection of secondary and tertiary structure in RNA[7]; here it is possible to pinpoint columns in multiple alignments that covary with much greater efficiency, due to greater constraints on ribonucleotide pairing. If such analysis could be brought to bear on protein multiple sequence alignments, it could aid the prediction of protein structure, providing specific constraints on fold topologies of the polypeptide chain. For the present, the columns in a multiple alignment that covary and are close to each other cannot be distinguished from those columns that covary and are far from each other. Future work will aim to enhance the distinction using information on protein secondary structure to filter out the "noise" contributed by these columns.

(5) References

(1) **Altschuh,D., Vernet,T., Moras, D., Nagai,K.** (1988)
Coordinated amino acid changes in homologous protein families.
Prot.Eng., 2:193-199.

(2) **Vernet, T., Tessier, D., Khouri, H.E., Altschuh, D.** (1992)
Correlation of coordinated amino acid changes at the two-domain interface of cysteine proteases with protein stability.
J.Mol.Biol., 224:461-471.

(3) **Taylor, W.R.** (1988) A flexible method to align large numbers of biological sequences.
J.Molec.Evol., 28:161-169.

(4) **Taylor, W.R., Jones, D.T.** (1993) Deriving an amino acid distance matrix.
J.Theor.Biol, in press.

(5) **Taylor, W.R.** (1986) The classification of amino acid conservation.
J.Theor.Biol., 119:205-218.

(6) **Taylor, W.R.** (1993) Compensating changes in protein multiple sequence alignments.
Prot.Eng.,submitted.

(7) **Winker,S., Overbeek,R., Woese,C.R., Olsen,G.J., Pfluger,N.** (1990)
Structure detection through automated covariance search.
Cabios 6 no.4:365 - 371.

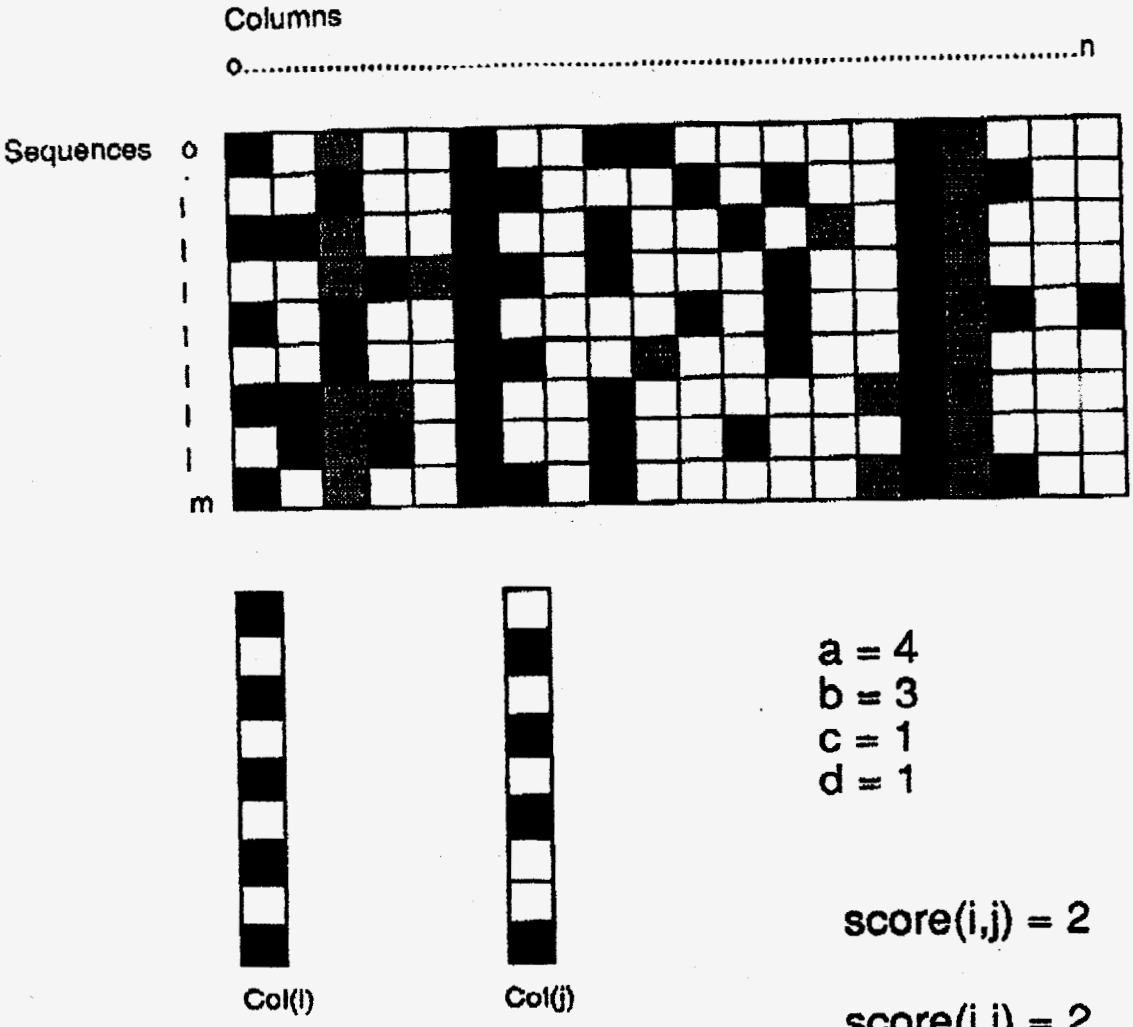
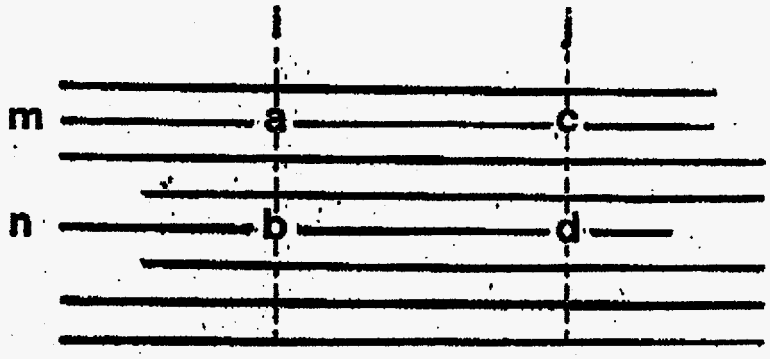

Columns
o.....n
Sequences
o
m
Col(i)
Col(j)
a = 4
b = 3
c = 1
d = 1
score(i,j) = 2
score(j,i) = 2
total_score(i,j) = 4

Table 1 :Various sample sizes (N) of the best conserved (cn) and best compensated (cr) pairs of residues were compared to the mean pairwise distance over all pairs in each protein. The accumulated differences in the means (A) are given in the columns headed cn-cr, al-cr, al-cn, while the columns headed cn< al, cr< al, cr< cn give the number of proteins in which one mean is less than the other.

N	cn-cr	al-cr	al-cn	cn<al	cr<al	cr<cn
10	-2.15	0.45	2.60	13	10	4
20	-1.70	0.74	2.44	13	11	5
30	-1.34	0.81	2.15	12	9	6
40	-1.62	0.72	2.34	12	9	5
50	-1.94	0.35	2.28	12	9	4

Figure 2 : In a multiple alignment of protein sequences, two positions i and j are considered. in method 2.2, each pair of sequences m and n is taken and the score *uijmn* is calculated.



A METHOD TO RECOGNIZE DISTANT REPEATS IN PROTEIN SEQUENCES.

Jaap Heringa and Patrick Argos

European Molecular Biology Laboratory (EMBL)
Meyerhofstrasse 1
Postfach 10.2209
D-69012 Heidelberg
Germany

An automated algorithm is presented that delineates protein sequence fragments which display similarity. The method incorporates a selection of a number of non-overlapping local sequence alignments with the highest similarity scores and a graph-theoretical approach to elucidate the consistent start- and end-points of the fragments comprising one or more ensembles of related subsequences. A multiple alignment of the resulting fragment ensemble(s) is performed. Finally, a profile is constructed from the multiple alignment to detect possible and more distant members within the sequence. The method tolerates mutations in the repeats as well as insertions and deletions. The sequence spans between the various repeats or repeat clusters may be of different lengths. The technique will be shown using a number of proteins where the repeating fragments are known based on information additional to the protein sequences.

Physical Mapping in the Presence of Errors: The Chimeric Clones Problem

David S. Greenberg[1], Sorin Istrail[1], and Michael Sipser[1,2]

(1) Sandia National Laboratories
Algorithms and Discrete Mathematics
Dept 1423
Albuquerque, NM 87185-5800

(2) Massachusetts Institute of Technology
Department of Mathematics
Cambridge, MA 02139

An important problem for the Human Genome Project is to develop robust software technology for the physical mapping of chromosomes. The process of creating a physical map is the divide/merge step in a divide-and-conquer approach to sequencing DNA and related problems. One large segment of DNA is reduced to several smaller pieces by physically breaking it apart. The process of breaking into pieces does not preserve information about the order of the pieces in the w hole and thus a combinatorial problem results to reorder the pieces. Various experimental errors complicate the problem. A major source of difficulty -- which seems to be inherent to the recombination technology -- is the presence of *chimeric* DNA clones. It is fairly common for two disjoint DNA pieces to form a chimera, i.e., a fusion of two pieces which appears as a single piece.

Attempts to order chimera will fail unless they are algorithmically divided into their constituent pieces. In an editorial in the October 1992 issue of Nature, Peter Little comments on the breakthrough of the first high resolutions physical maps using the cloned DNA technology and the difficulties associated with chimerism " At the risk of belittling a substantial achievement, there are still some serious drawbacks to these YAC maps. Some 40% (chromosome 21) and not more than 50-60% (Y chromosome) of the YACs contains artefactual hybrids of 21 or Y DNA with DNA from some other chromosome. These chimeric clones are very problematic to work with -- how do you know which piece of DNA comes from the correct genomic region ?" The chimeric clone problem has received only passing attention in the literature until now. In collaboration with Eric Lander of the Whitehead Institute we are

devising strategies for tackling this problem based on the optimization of several natural objective functions. Not too surprisingly, the computational complexity of such optimizations turns out to be invariably NP-complete (i.e., very likely intractable). Indeed, connections between the physical mapping and the Traveling Salesman Problem have been made repeatedly in the literature. To overcome this apparent computational intractability we have developed several algorithms including fast performance-guaranteed approximation algorithms. Through probabilistic analysis and simulations on synthetic data we are evaluating the performance of our algorithms. It turns our that one of our optimization functions is extremely successful in identifying chimeric clones and thereby allowing the creation of physical maps that relyably include almost the entire biologically correct one.

A first evaluation of our software library devoted to physical mapping of chimeric clones will be performed by researchers at the Whitehead Institute of Biomedical Research, and at Los Alamos Genomic Research Center.

(*) Supported in part by the U.S. Department of Energy under contract DE-AC04-76 DP00789

DETERMINATION OF ARCHITECTURAL ELEMENTS OF RNA STRUCTURE

Danielle A.M. Konings, and Robin R. Gutell

Department of Molecular, Cellular and Developmental Biology.
University of Colorado
Campus Box 234

RNA structure has been elucidated by a variety of computational and experimental methods. Comparative sequence analysis, as one of these methods, deduces a common secondary structure for RNAs belonging to a set of homologous sequences through correlation analysis based on a multiple sequence alignment. This method is responsible for the derivation of many higher-order structure models including tRNAs and 16S and 23S rRNAs. Underlying the complex architecture of these RNA structures is a repertoire of simpler structural elements. An understanding of these simple structural elements is needed to appreciate fully RNA structural and functional diversity. In addition, this knowledge is essential to further improve the accuracy of predicting RNA secondary structure from a single sequence based on thermodynamic or kinetic based principles. Starting from these complex higher-order structures as derived by comparative sequence analysis, we aim to identify their underlying structural constraints which will consist of simpler structural (and associated sequence) motifs. An example of such a structural motif are the specific sequence constraints that are associated with hairpin 'tetraloops' in rRNA.

Two research directives will be discussed that aim at increasing our understanding of RNA structure:

1. *Decipher RNA structural constraints.*

The objective here is to identify structural constraints of complex higher-order structures. Among others the analysis will be concerned with the distribution of specific RNA structure motifs (analogous to the tetraloop motif), position-specific motifs (e.g. protein binding motifs) and structural motifs that interchange at specific positions. We have developed a computational method that allows us to address these questions. The basic premise of comparative sequence analysis is that homologous sequence elements,

as defined by the sequence alignment, form equivalent units in the higher order structure of the RNA molecules. This principle allows us to utilize the multiple sequence alignment of a set of RNAs in conjunction with the secondary structure coordinates of a reference sequence to search for structural motifs throughout the entire set of sequences.

To allow for a systematic collection of structural constraints our searches for RNA structural elements are performed in a hierarchical manner: starting with the identification of generalized structural features and then progressively searching for more detailed and specific structural features. Our initial searches address general structural constraints such as the base composition of paired versus unpaired nucleotides, the distribution of base pair and consecutive base pair types in general and those associated with the closure of different loop types, and sequence constraints of loops according to the loop type and length. In general, however, the interest in specific searches evolves during the course of the analysis based on the specific outcomes of the successive steps. Thus far we have performed the statistical analysis on large collections of sequences of tRNA and 5S, 16S and 23S rRNAs. The analysis has revealed several structural features that are in agreement with known biophysical data as well as a number of unreported structural principles.

2. *Evaluation of free-energies and rules for RNA structure calculation.*

In parallel to the statistical analysis of sets of comparatively-derived structures as described above, we have started to identify structural elements that are presently poorly predicted by standard thermodynamic rules. Here we compared and contrast thermodynamically calculated structures with those identified by comparative sequence analysis. The elements identified by this analysis can serve subsequently as focal points in our statistical analyses or directly in the biophysical elucidation of improved folding principles.

A preliminary analysis of 16S rRNA structures of different phylogenetic groups in this way has revealed large differences in the predictive value of their respective structures. Whereas the thermodynamic structure prediction for eubacteria and archaebacteria is around 60% in terms of predicted base pairs, this percentage is only around 30% for mitochondria and eukaryotes. This imposes the question whether different phylogenetic classes of RNAs use

distinct structural elements or folding principles (e.g. the relative importance of proteins for stabilization) to generate their functional higher-order structure. To further address this question, a detailed mapping of the differences between the two types of structural models in terms of their association with basic structural elements, such as long-range helical elements, hairpinloops and multistemloops, will be performed.

COMPUTATIONAL MOLECULAR BIOLOGY IN LOGICAL PERSPECTIVE

A.K. Konopka

Particularly in Biology heuristic reasoning can be conclusive. Yet biologists face the problem of not being able to specify all the rules applied to derive conclusions. Nor are they able to list all assumptions on which those rules ought to operate. It seems that these inabilities are a reflection of the complexity of biological systems themselves. Biological phenomena are often represented by models that are still too complex to be described in a communicable manner. Further and further modeling is required until our observations can be conveyed in a linguistically comprehensive way. The cascade of models gives us the advantage of creating "communicable reality" but does not help us to judge the evidence pertinent to "real" (i.e., not necessarily communicable) reality. To the contrary, the more advanced a model in a cascade is, the further is its "distance" (in terms of number of modeling steps) from the modeled system.

From a logical point of view there are two problems here. First, we have no formal system of inference to judge formal correctness of observation sentences that have a variable true value (credibility) and that use ill-defined terms. Second, we have no formal system to judge the material adequacy of sentences derived from a "distant" model to the properties of modeled phenomena.

The problem of formal correctness seems to be solvable in principle. Progress in dealing with it can be noticed already in the fields of Artificial Intelligence (AI), Pattern Recognition (especially development of fuzzy mathematical techniques) and Situation Logic. In the near future we can expect to have formal tools to derive plausible conclusions from "imprecise" premises. Or, at least, we will have the option of relying on a machine (i.e., reliable AI software) that will perform plausible reasoning for us (I would probably exclude myself from the "us"). As far as formal judgement of material adequacy is concerned, it is unlikely that the problem is "addressable" in its generality. At least within mathematics the celebrated Gödel theorem precludes such a possibility. However, informal (more or less educated common-sense) judgements of material adequacy are possible and, as a matter of fact, all fields of science explore them.

Computational Molecular Biology (CMB) is a new science emerging from the liaison of computational technology and molecular biology proper (MBP). Its general goal is to understand biological phenomena through computational experiments and plausible reasoning. In a narrow sense, CMB is concerned with the informational (or symbolic) interpretations of biological phenomena that involve nucleic acid and protein sequences. The focus on symbolic interpretations distinguishes both CMB and MBP from physics, chemistry and other fields of "hard" science. However, the paradigms of CMB and MBP seem to have more aspects distinguishing them from each other than aspects in common. For one thing, instruments (say "measuring sticks") are different in CMB and MBP. Computer hardware and software along with plausible reasoning are examples of instruments used by computational biologists. Electrophoretic gels, centrifuges, spectrometers, probes of nucleic acids, antibodies and, perhaps, high individual tolerance to radiation are examples of equipment employed by MBP scientists. Because of the differences in research tools, both the data and data-associated percepts are different in CMB and MBP. So are the intellectual "folklores" that contribute to both paradigms. By inference, we ought to expect that criteria of material adequacy for observation sentences will not always be the same in CMB and MBP.

The presentation/discussion will (ideally) focus on protocols of pragmatic inference as seen (but not always clearly described) by most computational biologists. A general scheme of pragmatic inference for sequence research (part of CMB) is shown in Figure 1. The scheme is merely meant to illustrate the fact that vaguely defined "biological knowledge" is a vital factor allowing us to judge the material adequacy of outputs from each step of the "procedure". How exactly we should proceed to "correctly" implement our "biological knowledge" is currently unknown. Nor is it known how to clearly classify (or enumerate) this knowledge. Specific proposals of solutions for both those problems within sequence research will be described and (if time will allow it) discussed.

Suggested readings

1. Carnap, R. (1939). *Foundations of Logic and Mathematics*. Chicago Univ. Press, Chicago.

2. Gödel, K. (1931). *Über formal unentscheidbare Satze der Principia Mathematica und verwandter Systeme*. Monatshefte fur Mathematik und Physik 38,

173 - 198.

Konopka, A. K. (1993). *Sequences and Codes: Fundamentals of Biomolecular Cryptology*. In: Biocomputing: Genome Sequence Analysis, (D. Smith, Ed.). Academic Press, San Diego. in press.

Quine, W. V. (1980). *From a Logical Point of View*. Harvard University Press, Cambridge, MA.

Rosen, R. (1985). *Anticipatory Systems*. Pergamon Press, New York.

Rosen, R. (1991). *Life Itself*. Columbia University Press, New York

Shannon, C. E. (1949). *Communication Theory of Secrecy Systems*. Bell Syst. Tech. J. 28, 657 - 715.

Tarski, A. (1933). *The concept of Truth in Formalized Languages*. In: Logic, Semantics and Metamathematics, Ed.). Oxford University Press (1956), Oxford.

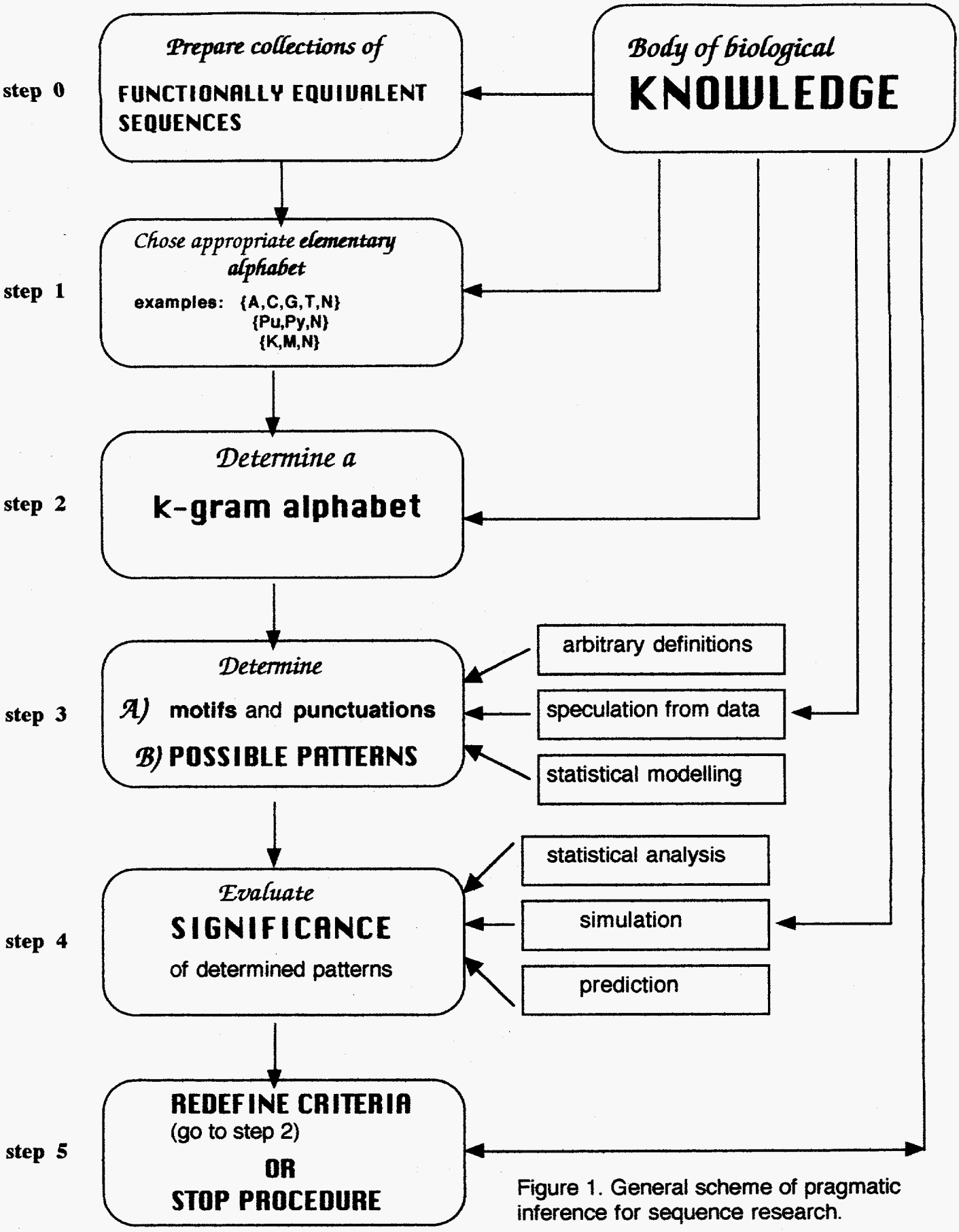


Figure 1. General scheme of pragmatic inference for sequence research.

Toward the Unification of Sequence and Structural Data for Identification of Structural and Functional Constraints

Chip Lawrence

Wadsworth Labs, NYS-DOH, Albany NY

&

NCBI, NIH, Bethesda MD

The identification and characterization of local residue patterns or conserved segments shared a set of biopolymers has provided a number of insights in molecular biology. Biopolymer sequences are observations from macro molecules that share common structural or functional features. The approach taken here rests on the notion that information may be most efficiently extracted from these observations through the use of a model that faithfully represents macro-molecular characteristics. Accordingly, our efforts are focused on statistical models which attempt to capture what we believe to be central features of protein structure, function, and change.

There are two major foci of our work in this area:

1) Threading of a sequence through structural motifs seeks to determine if a protein sequence fits a known protein structure, and in this way recognize folding motif. (Bryant and Lawrence, 1993)

2) Multiple sequence alignment via the Gibbs sampling algorithm seeks to identify position specific "free energy" models for residue sites in common "core elements" and simultaneously the alignment (Lawrence et al. 1993).

Because these two efforts address apparently different problems, they appear substantially unrelated. In fact, both are based on a common stochastic macro-molecular model. Five basic characteristics of protein structure, function, and change form the basis of these two collaborative efforts.

1) Proteins are stabilized by the energetic interactions among its residues, and the interactions of its residues with water, the peptide

backbone and ligands. These interactions determine a protein's structure and function.

2) When biopolymers have been subjected to a limited amount of evolutionary change, their commonality stems primarily from their mutational history. This case is of little interest to us. Rather, the focus here will be on the more difficult case that arises when the sequences have been subjected to extensive change, and any common patterns that remain are subtle. Among such distantly related sequences, common features stem primarily from structural or functional constraints. These constraints arise from the energetic interactions among residues or between residues and ligand. The relationship between energetic constraints and frequencies forms that basis of statistical mechanics, pioneering by Gibbs and Boltzmann. There is an analogous relationship for residue frequencies subject to random point mutations (Polh,1 971) (Berg & von Hipple 1987) (Bryant and Lawrence, 1991). This relationship suggests that the use of residue frequency models can be a valuable tool for representing the structural and functional constraints.

Given an alignment, the joint distribution of the residue types at the W positions in the common core may be represented by a multinomial residue frequency model. Interactions of residues with ligand, backbone atoms and water are essential to protein structure and function, and impose first order constraints on residue frequencies. Forces between pairs of residues are also key determinants of protein structure, and impose pairwise interaction constraints on residue frequencies. Since, the multinomial model is a member of the exponential family (Kendall and Stuart, 1952), and we consider at most pairwise interactions, the log joint distribution of residue frequencies may be described as a sum over first order terms plus the sum of pairwise terms over the set that mutually interact (Besag, 1974). In other words,

$$\log(P(R \mid A)) = \sum_{d=1..W} u_{d,r} + \sum_{C} u_{i,r,j,s} \quad (1)$$

where C is the set of residue pairs that make contact, and the $u_{d,r}$ and $u_{i,r,j,s}$ are respectively the first and second order free energy parameters.

3) Proteins or protein motifs that share a common structure share a common core. This core is composed primarily of ungapped segments of secondary structure interrupted by variable length loops.

4) The primary determinants of protein function are its energetic interactions with ligands. Thus, proteins or protein motifs that share a common function nearly always share common ligand based constraints. Ligand interactions often involve residues in a subset of a proteins loops. Since the geometry of these loops is tightly constrained to maintain these interactions, the lengths of these loops are nearly always preserved. Furthermore, residue frequencies in these loops are constrained by the energetics of their interactions with ligand.

5) Biopolymer sequences are misaligned by tranpositions, insertions/deletions, and gene duplications. In proteins these events result in variations in loop length. No direct data are available concerning the effects of these events in biopolymer sequences. Nevertheless, the probability of these events can be inferred from available sequence data.

In the 1970's it became widely recognized that many statistical problems are most easily addressed by pretending that critical missing data are available. In fact, for some problems, statistical inference is facilitated by creating a set of latent variables, none of whose values are observed (Goodman, 1974). The key observation was that conditional probabilities for the values of the missing data could be inferred, by application of Bayes theorem to the observed data. Statistical inference based on this concept was first described by Orchard and Woodbury (1972) and called the "missing information principle". Its application became widely known through a deterministic maximum likelihood algorithm, the expectation maximization (EM) algorithm (Dempster et al., 1977).

Geman & Geman (1984) developed a sampling based approach, which they named the Gibbs sampler. It was developed for the case in which the posterior distribution is a complicated, and thus difficult or impossible to obtain by direct integration. They employed this sampling algorithm both to develop a Bayesian description of the complete posterior distribution, and to find maximum a posteriori (MAP) estimates. They chose the name the Gibbs Sampler because a key required theorem from statistical physics , the Hammersley Clifford theorem, employs Gibbs/Boltzmann potentials to model joint probabilities from a complete set of conditionals. The use of sampling methods for problems involving missing data was first undertaken by Tanner & Wong (1987) and Li & Kim-Hung (1988). This sampling approach and its extensions have become a topic of great interest in statistics in the last few years (Gelfand & Smith, 1990),(Smith & Roberts, 1993). Most statistical

applications have little connection with statistical mechanics, thus the names Gibbs sampling has fallen into disfavor among some statisticians. Because the connections of this work with statistical physics name Gibbs Sampler is entirely appropriate.

The missing information principle was first used for sequence alignment to develop a block based (EM) algorithm for the identification and characterization of common motifs in biopolymer sequences (Lawrence and Reilly, 1990). This work subsequently was extended to permit small variations in the spacing of pairs of blocks (Cardon & Stormo, 1992). More recently, EM algorithms for gap-based alignment methods, in the form of Hidden Markov Models (HMM), have been described (Hausler et al., 1993). A more complete description of statistical aspects of the use of these ideas for misaligned data is given by Lawrence and Reilly (1992). Following this tradition, Bryant and Lawrence (1993) have recently presented a statistical model which imputes the alignment of a sequence to a structural motif, and Lawrence et al. (1993) have recently developed a Gibbs sampling algorithm which imputes the alignment of multiple sequences.

Threading:

In threading, the free energy parameters of equation (1) are taken as known. In fact, estimation of these parameters using the observed frequencies of residue pairs by distance in the protein data bank, was a major focus of the analysis conducted by Bryant and Lawrence (1993). With these known, the probability of an alignment is

$$P(A \mid R)) = \exp((\text{Sum}_{d=1..W} u_{d,r} + \text{Sum}_{C} u_{i,r,j,s} / Z)) , \quad (2)$$

where Z is the sum over all possible alignments. When all alignments the most probable can be identified from equation (2). When the number of alignments exceeds computing limits equation (2) can be employed as the basis of a sampling algorithm for the identification of the most probable alignment.

A Gibbs sampler for multiple sequence alignment:

For this problem the free energy parameters of equation (1) are unknown and are estimated from the available sequence data using an iterative sampling algorithm. The algorithm iterates between equation (1) and equation (2) with

the goal of simultaneously identifying the alignment and the unknown residue "free energy" parameters. In spite of the important contribution of pairwise interactions to protein stability, much of the information contained in protein motifs is captured by first order terms alone. Even when ligand specific effects are ignored, over 65% of the information concerning residue pairs frequencies in proteins of known structure is captured by first order "hydrophobicity" terms (Bryant & Lawrence, 1993). Furthermore, because the number of pairs is large, it is not clear that the additional energetic information from the pair terms will compensated for the substantially missing information penalty when structures are unknown. Accordingly, to date we will restrict attention to first order residue frequency models, for problems in multiple sequence alignment.

A more complete description of the mathematical models and algorithmic methods used will be given, and applications to subtle multiple sequence alignment problems and folding motif recognitions problems will be presented.

References:

Berg , O. G. and von Hippel, P. H. (1987), J. Mo. Biol. 193, 723.

Besag, J. (1974), J. Roy. Stat Soc: Ser B 35, 192-236.

Bryant, S.H. and Lawrence C.E. (1991) Proteins Struct. Func. Gene. 9, 108-119

Bryant, S. H. and Lawrence, C. E. (1993), Proteins Struct. Func. Genet. 16, 92-112.

Cardon, L. R. and Stormo, G. D. (1992), J. Mol. Biol. 223, 159.

Dempster, A. P.,Laird, N. M., Rubin, D. B. (1977), J. Roy. Stat. Soc.: Ser B 39, 1.

Gelfand, A. E. and Smith, A. F. M. (1990), J. Am. Statist. Assoc. 85, 389.

Geman, D. and Geman, D. (1984), IEEE Transaction in Pattern Analysis and Machine Intelligence 6, 721.

Goodman, L. A. (1974), Biometrika 61, 215-231.

Kendall M.K. and Sturat A., (1977) The Advanced Theory of Statstics, Vol I, Macmillian Pub. Co., NY.

Lawrence C.E., Alschul, S.F., Boguski, M.S., Liu, J.S., Neuwald, A. and Wootton, J.C., (1993) Submitted.

Lawrence, C. E. and Reilly, A. A. (1990), Proteins Struc. Func. Genet., 7, 41.

Lawrence, C. E. and Reilly, A. A. (1992), Biometrics Labortory Technical Report # 121.

Li and Kim-hung (1988) J. Statist. Computer & Simul. 57-59

Orchard, T. and Woodbury, M. A. (1972), Proceedings of the Sixth Berkeley Symposium on Mathematics, Statistics and Probability (University of California Press, Berkeley), Vol. 1, pp. 697-715.

Pohl , F. M. (1971), Nature New Biol 234, 277.

Smith, A. F. M. and Roberts, G. O. (1993), J. Roy. Stat. Soc.: Ser B 55, 3.

Tanner and Wong, (1987) JASA 805-811.

EVOLUTION ON RUGGED LANDSCAPES

Catherine Macken

Theoretical Division
Los Alamos National Laboratory
Los Alamos, NM 87545

We have studied walking on rugged fitness landscape as a model of the process of affinity maturation in an immune response. Affinity maturation appears to be the outcome of a mutation and selection process, described as "neo-Darwinist evolution in real time". Divorced from the biological context, walking on a fitness landscape applies to the appearance of a general optimization process.

Here, we outline a model of affinity maturation in order to introduce the formalism of fitness landscapes, and to develop an intuition about optimization in high dimensional space. Then we discuss more general issues, such as comparison of optimization techniques, and classification of landscapes.

Rugged fitness landscapes have been introduced in many areas of biology and physics to study, for example: affinity maturation of antibodies by somatic hypermutation, protein folding, RNA folding and evolution, species evolution, and spin glasses.

A fitness landscape consists of two components: sequence space, and a fitness function. **Sequence space**, S, is an abstract representation of the collection of all individuals of interest (protein, antibody, RNA molecule, glass states, etc.) as a set of strings or sequences of elements chosen from an appropriate alphabet. A **fitness function** assigns a real-valued "fitness" to each sequence in S. In principle, fitnesses can be plotted as heights of a landscape above the multidimensional sequence space. To model evolution on a rugged landscape, a rule is used to describe permissible moves through sequence space. Here we consider the rule in which only one element of a sequence can be changed in one unit of time. Thus, paths through sequence space involve moving between so-called **one-mutant neighbors.**

The nature of the landscape depends entirely on the fitness function. In most settings, this function is difficult to determine realistically, and therefore fitnesses are assigned in a somewhat arbitrary fashion. An extreme example, which nevertheless captures aspects of reality, is to assign a fitness value chosen at random from a probability distribution such as the

normal or uniform. Fitnesses are thus independent of sequence, and the resulting landscape has a characteristically rugged nature. Clearly, to assume that fitnesses are random functions of sequence is extreme. Site-directed mutagenesis reveals that some point mutations in a protein have little effect on its fitness, while others can drastically change fitness.

As an attempt to quantitatively describe the possibility of single mutations leading to small fitness changes, Kauffman introduced the NK model. In this model the fitness of a sequence of length N is a sum of fitnesses of the N individual elements of the sequence, but fitnesses of individual elements are assigned for each configuration of $0 \leq K \leq N - 1$ other elements. By altering the value of K, the ensuing landscape exhibits different degrees of ruggedness. The N K model has been studied in simulation experiments and to some extent analytically. Importantly, the assumption of additive fitnesses leads to strong central limit theorem effects in which all fitnesses cluster around the mean fitness with decreasing variability as N increases. Further, the ability to tune the degree of ruggedness by altering K results from the fact that random fitness samples of size 2^K have larger extrema as K is increased. We propose an alternative approach to a "tunably rugged" landscape that achieves variation in ruggedness by means of a model of physical structure within a sequence. It has the additional advantage of easy analysis and freedom from dominance by central limit theorem effects. We call our model the **block model** since it treats a sequence as a collection of independent blocks of elements. Fitness correlations among neighboring sequences are then a natural consequence of the block structure, and can be "tuned" by changing the number of blocks. The original motivation for our block model was the observation that molecular sequences often have natural partitions. When blocks of a sequence are long, the model is easy to study analytically: results of earlier theory (1) can be combined, using convolutions, into theory about the block model of sequences.

Consider a sequence of length N composed of B blocks. In considering antigen-antibody or protein-ligand interactions, we use the binding free energy (or log affinity) as a fitness. Thus, fitness of a sequence is $U = \sum_{i=1}^{B} U_i$, where block i has length n_i, $N = \sum_{i=1}^{B} n_i$, and U_i is the fitness of block i. We assume that blocks contribute independently to sequence fitness, and choose the block fitness, U_i, randomly from a continuous probability distribution G_i. Changing the number of blocks in the block model changes the characteristics of the fitness landscape. Intermediate levels of correlation between fitnesses of neighbors occur for $1 < B < N$.

Important results exist on the characteristics of a landscape, and of walks on a landscape. Many of these results are independent of the fitness distribution G, depending instead on a mechanism for assigning a rank to each possible sequence. When walks are restricted to movements in an uphill direction only, they quickly terminate at a local optimum. The length of walks increases with the degree of correlation between neighbors in sequence space.

The model for affinity maturation invokes a "random ascent" method of optimization, which may be compared with other optimizers, such as greedy algorithm, using, for example, reverse hill-climbing.

An important question is: How might landscapes be classified into categories within which optimization procedures have similar properties? To date, the correlation length of a landscape is the only such classification measure that has been proposed. However, a single statistic is unlikely to capture all of the influential properties of a landscape.

With the availability of a variety of optimization techniques, often incompletely understood, a correspondence between the characteristics of a landscape and preferred method of walking on the landscape is needed. Classification of landscapes is a first step toward establishing such a correspondence.

1. Macken, C. A., Hagan, P., and Perelson, A. S. (1991). Evolutionary walks on rugged landscapes. SIAM J. Appl. Math. 51, 799 - 827.

JUST-IN-TIME INFORMATION FORWARDING:
Biological Systems Viewed As Shannonian Communication Systems

Stephen A. Modena
Dept. of Crop Science -- Box 7620
North Carolina State University
Raleigh, NC 27695-7620
nmodena@unity.ncsu.edu

THE WAY THINGS ARE

"Originally, the flow of information involving the genetic material was termed the central dogma: DNA transferred information to RNA, which then directly controlled protein synthesis. DNA also controlled its own replication." [0] However a growing body of observational and experimental evidence has necessitated various addenda.

Molecular biology today seems to follow the Latin model: The Central Dogma is presented as the First Declension, regular in structure but not highly used. Subsequent declensions are introduced, replete with their structural irregularies, as the powerful main stays of cellular "life" at the molecular level. The final declension is a catch all for the exceptions not fitting well anywhere else and possibly not very important in the broader view.

Suppose one modernized the definition: "The central dogma is a description of the direction of information transfer among DNA, RNA and protein." [0] It still fails to accomodate the following: "The insertion of transposible elements into plant genes set the stage for a variety of interactions that can lead to alterations in normal transcriptional processes. The complex nature of these interactions...reflects the insertion of one intricate set of regulatory and processing signals into another." [1]

THE WAY THINGS COULD BE

I propose a successor model: Just-In-Time Information Fowarding. It will emphasize information, time, dynamics and noise; be scalable; be able to represent a biological work engine requiring energy dissipation; and most importantly, not be expressed explicitly in terms of specific cellular elements, unlike the tradtional central dogma. The "Just-In-Time Information Forwarding Model" is based on the Shannon General Communication System, as set forth by Shanno[2,3] with the specific augmentation by Rothstein[4], along with the addition of two heuristics needed for a genetics-driven biological system.

Any communication system has two main levels of abstraction: an end-to-end portion and a connectivity portion, illustrated in Figure 1, redrawn from [4]. The communication system exists to convey INFORMATION between the SOURCE and the DESTINATION in the presence of NOISE. The SOURCE and the DESTINATION are the end-to-end portion of the model. Let's defer specific identification of the source or the destination, except to say that the source is forwarding critical information through time and distance and hopefully it will arrive at the destination "just in time."

A SHANNONIAN SYSTEM

Shannon's Theory is a mathematical presentation, rooted in an ensemble of critical, interoperating components. The General Communication System is a logical machine operating through a physical machine. Briefly said, an INFORMATION SOURCE emits MESSAGES that a TRANSMITTER converts into physical SIGNALS, with encoding as needed; these are conveyed through a CHANNEL, i.e. the real world with time, distance and energy attributes; during CHANNEL transit, the SIGNALS may be altered by combining with pervasive NOISE always found in the CHANNEL; possibly altered SIGNALS arrive at a RECEIVER, which makes imperfect decisions while converting the SIGNALS to equivalent MESSAGES, employing whatever decoding appropriate; these MESSAGES hopefully are understood and have utility to the DESTINATION.

THE STRUGGLE BETWEEN THE LIGHT AND THE DARKNESS

Information and noise are the main facets of Shannon's model. In his worst case analyses, Shannon assumed "white noise," because of its natural and convenient maximum entropy property. Rothstein[4] remapped Shannon's model into a general model of a measuring-procedure-and-apparatus (for example, the case of an analytical balance). To account for the precision and accuracy biases inherent in fluctuating measured-values, he renamed NOISE to "error source" (Fig. 1). Rothstein's symbolic remapping has numerous ramifications and applied consequences.

I am remapping the Rothstein-modified Shannon Model into the biological sphere, which ought to be legitimate since Shannon himself precedes me[5]. This model is proposed from a particular point-of-view: that we, experimental and computational biologists, are engaged in the reverse engineering of the machinery of life. As such, we often defer the larger questions of "who" and "why" in favor of concentrating on the more immediately tractable aspects of our favorite system. I would like to emphasize that I am borrowing the Shannon system of ideas in its entirety, rather than selecting one or two features and discarding the remainder--a common practice which caused Shannon himself both distress and disillusionment[6].

JUST IN TIME DELIVERY OF CRITICAL INFORMATION

My selection of the "just in time" imagery, the first heuristic, arises from a specific example of RNA editing in the trypanosome mitochonrion. An "AUG" start-signal for polypeptide synthesis is not found in the genomic DNA cistron and is not present in the transcribed mRNA. "Just in time," a correctly placed "AUG" is edited into the mRNA[7], allowing the appropriate polypeptide to be made by the ribosomal complex. In effect, I have chosen a DESTINATION, the just-in-time correct initiation of a vital polypeptide. The MESSAGE can be taken to be the fully complete mRNA, containing among other features the "AUG" nucleotide triplet in correct junxta-position to "other" signals needed by the ribosomal complex.

Under the influence of the central dogma, one author wrote this wrote this about RNA editing: "While other forms of RNA processing maintain primary sequence correspondance between gene and transcript, RNA editing disrupts this informational linkage by altering the actual sequence of an RNA molecule after it has been transcribed."[8] It is factually correct, but conceptually perverted.

The "Just In Time" heuristic forces one to clearly define the DESTINATION and the probable ensemble of messages that might be received via the communication system.

INFORMATION FORWARDING

Information forwarding is the second heuristic. If the mRNA mentioned above is not constitutively expressed, then one must account for it's timely appearance. Symbolically, one can say that this mRNA, positioned on the ribosomal complex just about to trigger protein synthesis, must be reincarnated from time-to-time by means of an information forwarding scheme. In essence, the source and the destination are the "same" temporal phenomenon or entity with an intervening existence as pure "information." A similar notion is embodied in the biological concept named "alternation of generations.

The "Information Forwarding" heuristic forces one to identify the SOURCE and the probable ensemble of messages that might be transmitted via the communications system.

Identifying the sent and received message ensembles forces a thoughful specification of the communication channel(s) used; any serialization or parallelization, i.e. scaling; timing requirements or constraints; and any encoding-decoding parameterizations and specialisations.

NOISE AND ERROR ARE TWINS

In a thermodynamic Universe, noise is pervasive and inescapable. During the conveyance of information through time and distance as a channel signal, noise interacts with that signal in an algebraically additive fashion[2]. The net result is that a certain potion of the original information can become irrecoverable, because it has been converted to error. Noise imposes a natural upper limit to system efficiency. Noise has another consequential effect: it quantizes continuously variable information. Quantization and error minimumization can both be dealt with via codes and encoding schemes.

LEARNING FROM REAL SHANNONIAN SYSTEMS

Research in radio spread spectrum systems shows that the statistical structure of noise sources has a profound effect on the success and efficiency of any particular system of encoding in a communications system.[9]

Often a communication channel must be shared among multiple signals. Because physical communication systems are finite machines, they have a capacity limi. When the capacity requirements of mutiple friendly signals sum to more than the available channel capacity, each signal begins to experience degradation because the other co-resident signals shift to operating as noise/error sources in a mutual fashion.

As experience with radio spread spectrum techniques has demonstrated, some sources of channel "noise" are actually co-resident signals designed to reduce channel efficiency or to direct well structured errors into other channel coresident signals.

ACTIVELY ADAPT OR DIE

Adaptive stategies have been embraced to counteract or tolerate channel capacity depletion by either friendly or enemy channel co-occupants. I believe that one of hallmarks of genetic systems is the evolutionary adaptive capacity to tolerate or coexist with friendly and unfriendly channel co-residents under competition for channel capacity resources. Some of the main strategies appear to be switching coding schemes; adopting better encoding to effectively increase channel capacity or create channel diversity; increased segmentation of the channel; resynchronization via time diversity to take adavantage of episodic fluctuations in channel capacity availability; and restructuring signals to take advantage of mutual entropy components of co-resident channel signals.

WORKED PROBLEMS

Several common molecular biology phenomena will be presented for remapping to the "Just-In-Time Information Forwarding Model." Hopefully, frank assessment and reassessment of ideas will result. :^)

[0] R.H. Tamarin, "Principles of Genetics", PWS Publishers, Boston, 1986 .

[1] C.F. Weil and S.R. Wessler, "The Effects of Plant Transposable Element Insertion on Transcription Initiation and RNA Processing", Ann. Rev. of Plant Physiol. Plant Mol. Bio. 41:527-52 1990.

[2] C.E. Shannon, "AMathematical Theory of Communication", Sys. Tech. J. 27:379-423,623-656 1948

[3] C.E. Shannon, "Communication in the Presence of Noise", Proc. Inst. of Radio Engineers 37:10-21 1949

[4] J. Rothstein, "Information, Measurement and Quantum Mechanics", Science 114:171-175 1951

[5] C.E. Shannon, "Prediction and Entropy of Printed English", Bell Sys. Tech. J. 30:50-64 1950

[6] C.E. Shannon, "The Bandwagon (Editorial)", I.R.E. Trans. Info. Theory IT-2:3 1956

[7] K. Stuart, "RNA Editing in Trypanosomatid Mitochondria", Ann. Rev. Microbiol. 45:327-44 1991.

[8] M.W. Gray et. al., "Transcription, Processing and Editing in Plant Mitochondria", Ann. Rev. Plant Physiol. Plant Mol. Biol. 43:145-175 1992.

[9] R.C. Dixon, "Spread Spectrum Systems", John Wiley & Sons, New York, 1984.

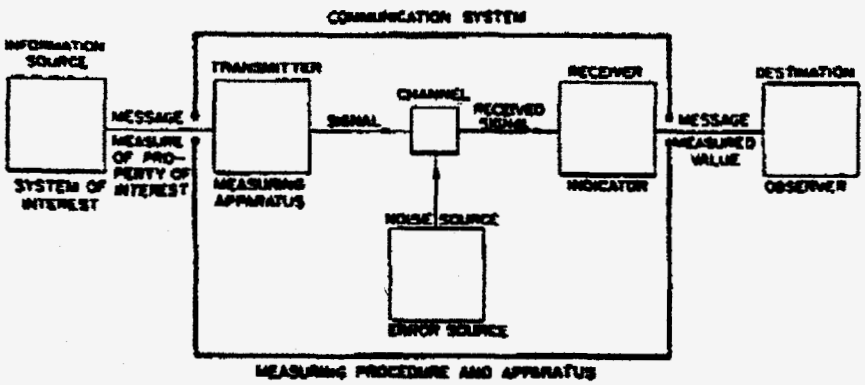


FIG. 1.

Towards DNA Sequencing by Hybridization

Pavel A. Pevzner

Departments of Computer Science
The Pennsylvania State University
University Park, PA 16802

In view of the limitations of current DNA sequencing technology, it would be advantageous to have a method for sequencing DNA that (1) does not require gel electrophoretic separation, (2) provides the sequence of a very long DNA fragments in a single experiment, (3) is amenable to automation.

Sequencing by Hybridization (SBH) is a challenging alternative to the classical DNA sequencing methods. The basic approach is to build an array (Sequencing Chip) of short oligonucleotides, to use hybridization for finding oligonucleotide content of an unknown DNA fragment and to reconstruct the original fragment by a combinatorial algorithm. Three major breakthroughs in SBH have reported recently; Southern et al. built the first sequencing chip, Drmanac et al. read the first 100 bp by SBH, Fodor et al. developed photolithographic technique for building sequencing chips with millions of oligonucleotides. It is becoming clear that SBH is a challenging approach and may be the most promising alternative sequencing method and that large-scale SBH projects are about to be launched. Even today a chip for sequencing hundreds to thousands of nucleotides might cost from a few dollars to tens of dollars when made by mass production.

The implementation of SBH requires the application of biochemistry, computer science and high technology. Recent studies indicate that the original SBH chip C(8) containing all 65,536 octanucleotides is insufficient for sequencing long DNA fragments. In particular, Pevzner demonstrated that even in the case of an ideal SBH experiment (no hybridization errors) one can hope to reconstitute a 200 nucleotide long sequence only in 94 out of 100 cases. This discouraging result indicates that additional joint efforts by biologists and computer scientists are required to make SBH practical. Sometimes biologists can not estimate the computer science limitations of the proposed SBH experiments. On the other hand computer scientists frequently

cannot anticipate what information useful for DNA sequence reconstruction can be obtained by modern biochemical techniques. We are trying to narrow this gap by presenting the links between the biochemical, computer science and technological aspects of SBH.

We also consider the problem of designing high-resolution chips for SBH. Recently Pevzner et al.,1991 demonstrated that the classical C(k) sequencing chips containing all k-tuples are very redundant and therefore inefficient. I present a new chip design which allows significant chip miniaturization without reducing their resolving power. This is a joint work with Robert Lipshutz.

ON THE SCHRÖDINGER QUESTION

Robert Rosen

Dept. of Physiology & Biophysics
Dalhousie University

ABSTRACT

Almost fifty years ago, the eminent physicist Erwin Schrödinger published a famous essay, entitled "What is Life?". We undertake a comprehensive re-evaluation of this basic question, both as Schrödinger himself envisioned it, and in the light of a half-century of further experience since his time, not only in biology, but also in physics, in mathematics, and in the theory of systems.

Today, Schrödinger's Essay has long since been embalmed as a classic embodiment of molecular-biological orthodoxy. We assert that, to the contrary, Schrödinger's arguments were then, and remain today, highly heterodox; indeed, quite incompatible with the dogmas which have evolved since then. We will attempt to reformulate these original arguments, and compare them to the distorted and truncated versions which have survived in the molecular biology of today.

We start from Schrödinger's original question, "What is Life?". To what extent is this even a legitimate scientific question? That is: to what extent can "life", the subject-matter of biology, be considered a *thing;* a legitimate object of scientific scrutiny in itself? If so, what kind of "thing" is it? Schrödinger clearly thought it was a thing in itself; not just a qualifying adjective for other things. Today's molecular biology would consider the question itself nonsensical.

Besides simply taking this question seriously, Schrödinger repeatedly insisted in his essay that, in his words, "new physics" was required to address it properly. On the other hand, people like Jacques Monod, writing decades later, denounced any such suggestion as "vitalism", and as incompatible with the "objectivity" of science itself.

Schrödinger's essay seems to devolve around an earlier argument of

his colleague Max Delbrück, which asserts that the "Mendelian gene", recognized and characterized in terms of phenotypic effects, had to be a molecule. Under this camouflage, Schrödinger actually concerned himself more with a converse question: when could a molecule be a Mendelian gene? The ramifications of this converse question were what led to the conviction that "new physics" was essential; at the same time, they led to the most familiar, yet the most thoroughly misunderstood parts of the Essay: the "aperiodic solid", the idea of a cryptographic relation between genotype and phenotype, and the idea of "feeding on order (negentropy)".

In modern terminology, Schrödinger is asking us to contemplate the stability of genomically forced, thermodynamically open material systems. Beyond a few vague mathematical hints, and a few heuristic physical rules of thumb, no part of the program Schrödinger was hinting at has even been approached. None of this program is addressed at all by the reductionistic character of modern molecular biology, nor have they yet been approached from conventional physical directions; indeed, it turns out that the preferred starting-point of conventional physics, the closed system, is so degenerate (structurally unstable) that how you open it is an infinitely more important determinant of how it will behave than what it is like when closed.

Schrödinger concluded his Essay by completely discounting the "machine analogies", going back to Descartes and before, which seemed to provide the backbone for any strategy to relate biology and physics. On the other hand, these analogies are precisely what have been retained in today's molecular biology; they comprise whatever vision and philosophy it has. Yet the very deficiencies of these machine metaphors go a very long way in specifying the shape of the vision Schrödinger was actually advocating.

Biopolymer Sequences and Structures

Peter Schuster
Institut fur Molekulare Biotechnologie
Jena
Germany

Properties and functions of biopolymers are considered as being the result of two classes of mappings. The first class consists of maps which assign a (three-dimensional) structure to every sequence. They are tantamount to mappings from genotypes into phenotypes. The second class of mappings deal with relations between structures and functions as expressed by free energies, activation energies, or other scalar quantities. They map structures into the real numbers. A landscape is understood as a combination of one map of each of the two classes. We thus have, for example, free energy landscapes, activation energy landscapes, fitness landscapes, etc.

RNA secondary structures are chosen as an example because only in this case we have sufficiently fast folding algorithms at hand which allow to handle millions of sequences and structures. The relation between RNA sequences and structures is considered as a mapping from one metric space of sequences into another metric space of secondary structures, called shape space. In sequence space the Hamming metric is used. An appropriate metric in shape space is obtained by tree editing. The secondary structures are converted into equivalent trees for that purpose. Several properties of RNA shape space will be discussed. In particular three major results are presented:
(1) we have many more sequences than structures,
(2) sequences folding into the same structure are randomly distributed in sequence space, and
(3) any random sequence is sourrounded by a ball in sequence space which contains sequences folding into all common structures.
The radius in Hamming distance of this ball is much smaller than the chain lengths of the sequences.

The meaning of these results for biological evolution and evolutionary biotechnology is discussed. Possible generalizations to real three dimensional structures of RNA molecules and to protein structures will be considered.

Evolution on Fitness Landscapes

Peter F. Stadler

Institut fur Theoretische Chemie der Universitat Wien
Wahringer Strasse 17
A-1090 Vienna
Austria

Evolution can be viewed as an adaptation process on a 'fitness' landscapes (at least in some systems). The dynamics of evolution are hence tightly linked to the structure of the underlying landscape. Global features of landscapes can be described by statistical measures like number of optima, lengths of walks, and correlation functions.

Statistical characteristics of RNA landscapes are accessible on the level of secondary structures. It turns out that RNA landscapes belong to the same class as well known optimization problems and simple spin glass models.

The evolution of a quasispecies on such landscapes exhibits three dynamical regimes depending on the replication fidelity: Above the localization threshold the population is centered around a (local) optimum, between localization and ``dispersion threshold'' the population is still centered around a consensus sequence, which, however, is non constant in time. For very large mutation rate the population spreads gas-like in sequence space.

ABSTRACT for "Open Problems in Computational Molecular Biology"
Telluride, July 1993.

NONALIGNABILITY AND NONCOMPACTNESS MODELS AND REALITY OF LOW-COMPLEXITY PROTEIN SEGMENTS

JOHN C. WOOTTON

National Center for Biotechnology Information, Building 38A, Room 8N805, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, U.S.A. E-mail: wootton@ncbi.nlm.nih.gov

Introduction

How much of amino acid sequence space is represented in natural proteins? Global computational analyses of sequence databases have recently shown surprising biases. More than half of deduced protein sequences contain at least one segment of low compositional complexity, consisting predominantly of one or a few amino acids, and such strongly biassed, interspersed segments comprise 15% of the residues in the database (Wootton & Federhen, 1993). Conserved sequence families, of the type common to a wide range of organisms and familiar since the 1950s from sequence alignments, may number fewer than 1000 according to some estimated extrapolations (Green et al., 1993), and such conserved domains and motifs may eventually account for less than half of the residues in natural proteins. Moreover, from a rapidly growing number of cases of random cDNA or genomic coding DNA sequenced without regard for function, deduced amino acid sequences show a non-random excess of long "medium-complexity" sequences in addition to the near-homopolymeric low-complexity segments.

Taken together, these analyses reveal a disturbing level of ignorance about protein structure, dynamics, interactions and evolution. The wealth of knowledge of relatively compact globular proteins, as derived in atomic detail from crystal and NMR structures, and as also represented by numerous alignments of conserved domains or motifs from multiple sequences, may provide a paradigm for understanding as little as half of natural amino acid sequences or subsequences. Low and medium-complexity segments, as far as can be inferred from the limited evidence available, generally evolve rapidly, rarely have unique conformations and show more flexibility in conformational dynamics and interactions with other macromolecules and solvent. New models will be

explored in this report for conceptualizing the physical properties of such "non-globular" sequences, and methods will be evaluated for their utility in analyzing and organizing the copious mass of low-complexity amino acid sequence data now available.

Nonalignability

The great majority of low-complexity sequences can not be sensibly aligned by conventional sequence comparison methods based on residue position. Their heterogeneous mixture of compositional biases confounds the statistical basis of most sequence alignment algorithms. More fundamentally, traces of mutational pathways of evolutionary descent are lost through insertions, deletions, substitutions and sequence repeats, except in comparisons of equivalent sequences in close organisms. Even corresponding low-complexity sequences from distant species, which are evidently similar in function and location, do not generally align in a unique manner by residue position. Mutational dynamics involving DNA replication slippage and repeat expansion may generate similar simple sequences de novo.

These properties have prompted an analysis of the premises of multiple sequence alignment methodology. Such alignments have relevant fundamental limitations when considered as representations (or "observations") of unknown common folds of residue chains. Some issues arise from discrete sybolization of a spatial continuum and arbitrariness of gaps. When multiple alignments are transformed into numerical models, additional problems arise from the data-dependence of the generalizability of the models, from effects of fixed dimensionality, and from the paucity of reliable independent evaluation criteria.

Classification of low-complexity amino acid sequences, and the possible detection of any recurrences of very subtle compositional or sequential patterns, may be based on unaligned sequences in the absence of prior knowledge of significant subsequences. Different pseudometrics have been evaluated for their ability to neighbor these segments by single and k-word composition, together with residue correlations within segments and definitions of local compositional complexity.

Figure 1 illustrates the results of one of these methods used to search the Swissprot database. The query sequence is a methionine-proline-glycine rich

segment of a type occurring in some small nuclear ribonucleoproteins and heat shock protein 70. The complete list contains approximately 100 examples of this class of sequence, which is unusually conserved in subsequence characteristics in the above proteins, is an important epitope in some autoimmune diseases, but is of unknown structure, dynamics and interactions. This search reveals that segments of the same compositional properties also occur in many other DNA binding and developmental control proteins from a wide range of organisms. Such methods may be extended to give a relatively comprehensive classification of the low-complexity sequences of the database. This process is complementary to the neighboring by Blast MSP scores, as used for the NCBI Entrez datasets, for which purpose the low-complexity sequences are filtered from the database prior to Blast neighboring.

One limitation arises in the question of what constitutes a class or family of low-complexity sequences. In many cases, in the absence of direct evidence for structure or dynamics, this has to be decided using purely statistical criteria. Relatively indirect empirical evidence is available in some cases from mutational effects or the location of segments in relation to conserved globular domains. Such cases, together with regularities that emerge from the unbiased classification of the segments, give a few general insights into the nature of sequence variation in functionally significant low-complexity sequences.

Noncompactness

To what extent do simple amino acid sequences imply noncompact, nonunique, flexibly dynamic 3-dimensional structures? This question has been explored using some simplified physical models and a comparison of different definitions of conformational entropy and structural complexity. The results give some insights into the sequence requirements for compact collapsed structures and also into the nature of molecular assemblies that exhibit larger-scale emergent "informed" thermal and dynamic properties.

References

Green P, Lipman D, Hillier L, Waterston R, States D & Claverie J-M (1993) Ancient conserved regions in new gene sequences and the protein databases. Science 259: 1711-1716.

Wootton J C & Federhen S (1993) Statistics of local complexity in amino acid sequences and sequence databases. Computers & Chemistry 17: (in the press).

```
QUERY SEQUENCE:        pgmmpaphmggppmmpmmgppppgmmpvgpapgmrppmgghmpmmpgppm
                       mrpparpmmvp

   NAME,
 RESIDUES    DISTANCE                SEQUENCE
 --------    --------                --------

1 SM11_RAT
   152-212      7.73  ppgrggppppmgrgapppgmmgpppgmrppmgppmgippgrgtpmgmppp
                      gmrppppgmrg

2 HS70_TRYCR
   616-666      8.38  gmagagmpggmpggmpggmpggmpggmpggmpggmpggmpggmpggmpgg
                      a

3 RSMB_HUMAN
   169-240      9.22  ppgrggppppmgrgapppgmmgpppgmrppmgppmgippgrgtpmgmppp
                      gmrppppgmrgppppgmrpprp

4 RU1C_HUMAN
    70-152     11.19  pppagamippppslpgpprpgmmpaphmggppmmpmmgppppgmmpvgpa
                      pgmrppmgghmpmmpgppmmrpparpmmvptrp

5 K10_DROME
   224-281     11.71  rgpppnrppprlmmgppmgpmgpgprgpgpmgpggpypgmpfpppvpgmr
                      gpghhgpm

6 SRP5_YEAST
   495-537     11.74  glmqnmmnmfgggmgggmgggmpdmnemmkmmqdpqmqqmakq

7 MAM_DROME
  1092-1156    13.84  vgvgvgvgvgvgvgvgnggvvggpgsggpnngamnqmggpmggmpgmqmg
                      gpmnpmqmnpnaagp

8 HS70_BRELC
   615-659     14.43  aaagaaggmpggmpgmpggmpgmpggmpgmpggmpggmpggmpgg

9 K2C2_XENLA
   299-374     15.30  ggaaggmgggggmgggmgggmgmgggmgmgggmgmgggmgmgggmgggmg
                      mgggmgmgggmgmgdgmgsghgggyg

10 HS70_PLAFA
   627-675     16.12  aagaaggmpggmpggmpggmpsgmpggmnfpggmpgagmpgnapagsgp

11 DC11_DROME
   752-794     20.36  qaqqmaqqnqmmmqqmqqrqwsedqaqiqqqqrqmmqqtpmmm
```

Figure 1. Neighboring a low-complexity amino acid sequence. The list shows the few nearest neighbors, amongst the low-complexity sequences of the Swissprot database, to the query sequence, which is a methionine-proline-glycine-rich subsequence of the human U1 SnRNP-associated protein. The hits are ranked in order of distance, as measured by an algorithm based on the pairwise residue covariances observed in the low-complexity subset of Swissprot.