Molecular Bioinformatics

Dan Gusfield, UC Davis
Tom Lengauer, GMD St. Augustin
Chris Sander, EMBL Heidelberg /ECI Cambridge

Dagstuhl-Seminar-Report, July 9 – 14 1995

Dagstuhl Seminar on Molecular Bioinformatics

The seminar was a sequel to the first Dagstuhl Seminar on Molecular Bioinformatics which took place from Sep 7 to Sep 11, 1992 and brought together computer scientists and applied mathematicians with biochemists and molecular biologists in order to discuss possibilities of cooperation in the field of computer-aided design of biomolecular sequences and structures. Since that time, several developments have taken place in Germany and internationally.

In Germany, BMBF (the Science Ministry) has funded a program on Molecular Bioinformatics. Within this program, eight large interdisciplinary projects are working on topics such as protein structure prediction, biomolecular docking, genomic databases, and RNA secondary and tertiary structure prediction. Internationally, efforts in computer science applications in molecular biology have increased significantly. Countries such as U. S. and Japan witness lively activity in these fields, albeit somewhat more concentrated on the genome itself than on genomic products such as proteins. Interaction between computer scientists and molecular biologists is intensifying, and a growing number of results are obtained with involvement of computer scientists.

The topic of the seminar included research on both the genome and its products, with somewhat of a emphasis on genomic products and on the relationship between those two kinds of research than on the genome itself. As the previous seminar, this seminar concentrated on algorithmic issues and touch issues such as data handling technology, computer graphics etc. only insofar as they relate to algorithmic issues.

In addition to the presentations, there was an evening tutorial on protein structures, a round of discussion on traditional and network-based teaching forms for bioinformatics, and a software demonstration.

The general feelings of the participants were that after the ground-breaking character of the first seminar, this seminar was marked by an increased understanding and depth of communication between the CS and the molecular biology side. It was clearly evident that the two communities came to understand each other and began to develop a common feeling of identity.

The facilities and procedures at Dagstuhl as well as the unique concept of the Dagstuhl seminars were praised by many participants. The participants expressed the hope that this seminar be succeeded with another seminar on bioinformatics in due time. While a few of the participants expressed the wish to have future workshops more strongly focussed, say, just on alignment of sequences and structures, the majority of the participants welcomed the broadness of the workshop and felt that this width of scope should be maintained of some more time.

Program

Monday, July 10

Tom Lengauer	Welcome
Morning Session	Chair: **Tom Lengauer**
John Kececioglu	Genome Rearrangements
Gene Myers	Developments in DNA Sequencing
Dan Gusfield	Parametric Alignment
Afternoon Session	Chair: **Dan Gusfield**
Ralf Zimmer	Parametric Alignment
Kevin Karplus	Using Simple Markov Models
Tandy Warnow	Reconsidering the Construction of Evolutionary Trees
Steve Altschul	Aspects of the Statistics of Local Sequence Comparison
Andreas Dress	Visualization Procedures Related to Abstract Similarity Analysis, Quasi-Crystals and the Traveling Salesman Problem, Part I: Pictures
Evening Session	Chair: **Chris Sander**
Chris Sander	Tutorial on Protein Structures

Tuesday, July 11

Morning Session	Chair: **Steve Bryant**
Andreas Dress	Visualization Procedures Related to Abstract Similarity Analysis, Quasi-Crystals and the Traveling Salesman Problem, Part II: Proofs
Volker Brendel	Application of Scoring in Sequence Comparison for Evolutionary Studies and Motif Recognition
Dalit Naor	Amino-Acid Pair Interchanges at Spatially Conserved Locations
Joachim Selbig	Clustering of Protein Structures on the Basis of Hexapeptides
Georg Casari	Specific Functional Regions in Protein Families
Afternoon Session	Chair: **Chris Sander**
Russ Altman	Probabilistic Representations and Algorithms for Analysis of Structure
Liisa Holm	Protein Structure Database Searches
Steve Bryant	Finding the Most Surprising Structural Similarities
Ralf Thiele	Threading by Recursive Dynamic Programming
Steve Bryant	An Alignment Model and Fast Algorithms for Protein Threading
Evening Session	Chair: **Robert Giegerich**
Giegerich, Shamir	Teaching Computational Biology
Altman, Karplus	
Vingron, Myers	

Wednesday, July 12

Morning Session	Chair: **John Kececioglu**
Mike Steel	Generating New Phylogenetic Information from a Collection of Trees
Ron Shamir	Physical Maps and Interval Graphs
Michael Gribskov	An Evolutionary Mixture Model for Describing Protein Sequence Families
Ina Koch	An Algorithm for Finding All Maximal Common Substructures and Its Application to Protein Structures

Thursday, July 13

Morning Session	Chair: **Martin Vingron**
Norbert Blum	On the Prediction of RNA Secondary Structure
Russ Altman	Constraint Satisfaction Methods for Modeling the Complete 30S Ribosomal Subunit and its Interaction with Transfer DNA
Robert Giegerich	Representing Large RNA Folding Landscapes
Dannie Durand	Ultraselfish Genes in Mice
Hans-Peter Lenhof	Protein-Protein Docking
Afternoon Session	Chair: **Dalit Naor**
Matthias Rarey	Time-Efficient Docking of Flexible Ligands into Active Sites of Proteins
Thomas Seidl	A Database System for Protein Docking
John Kececioglu	Multiple Sequence Alignment
Gene Myers	Rapid Similarity Search
Satoru Miyano	BONSAI Garden: Parallel Knowledge Discovery System for Sequences
Matthias Rarey	Demonstration: Docking

Friday, July 14

Morning Session	Chair: **Tandy Warnow**
Thure Etzold	Integration of Biological Flat File Databanks into an Object-Oriented Retrieval System
Jürgen Kleffe	Loglinear Models for Splice Site Recognition
Martin Vingron	Sequence Alignment and Phylogeny
Dan Gusfield	How to Think About Repetitive Structures
Tom Lengauer	Close and Goodbye

Algorithms for Evolutionary Distances between Genomes with Translocation

John Kececiolgu, University of Georgia, USA

A new and largely unexplored area of computational biology is combinatorial algorithms for genome rearrangement. Rearrangement mechanisms include inversion, transposition, duplication, and translocation, and a basic problem is to determine the minimum number of such events to transform one genome to another. This number is called the rearrangement distance between the two genomes.
We begin the algorithmic study of genome rearrangement by translocation. A translocation exchanges material at the ends of two chromosomes in a genome, which we model as an exchange of prefixes and suffixes of strings, where each string represents a sequence of unique markers along a chromosome. For the general problem of determining the translocation distance between two such sets of strings, we present a 2–approximation algorithm, and for a theoretical model in which the exchanged substrings are of equal length, we present an optimal algorithm. For genomes that have evolved by both translocation and inversion, we show there is a simple 2-approximation algorithm when the orientation of markers is not known, and a 3/2–approximation algorithm when orientation is known.
This is joint work with R. Ravi.

Recent Developments in Shotgun DNA Sequencing

Gene Myers, University of Arizona, USA

We begin with an overview of shotgun DNA sequencing and the computational characteristics of the problem — (1) unknown orientation, (2) incomplete coverage, (3) error in fragment reads, (4) repeats in the underlying target sequence, and (5) auxiliary constraints reflecting the particular experimental protocol used to produce the data. We emphasize and illustrate the increasing importance of the problems of negative sequence and accommodating constraints. We then briefly describe the decomposition of the computational problem into 3 phases:

(A) Overlap – find all pains of approximate overlap between fragments,
(B) Layout – determine an arrangement of the fragments (based on the overlap),
(C) Consensus – multi-align the fragments in regions above coverage is 3–or–move,

and recast the problem in graph–theoretic terms. Finally we give a series of graph reduction transformations that in practice *collapse* the graph almost completely without changing the objective of layout or precluding the optimum answer. This reduction is an essential first step towards correctly handling repeats and constraints.

Efficient Parametric Sequence Alignment with XPARAL

Dan Gusfield, University of California, Davis, USA

Most existing sequence alignment methods require the user to specify the value of specific parameters such as the value of each match and the penalty for mismatches, spaces and gaps. It is widely noted that the quality of the alignment is effected by the choice of parameter settings.
Parametric alignment solves the alignment problem as a function of variable parameters and decomposes the parameter space into maximal regions where an alignment is optimal. Our program XPARAL computes this decomposition and allows one to explore the parameter space by examing the co–optimals in this polygon. In this talk we review the status of XPARAL, see how it is applied to a problem of aligning secondary structure in proteins, and show that for most alignment models, XPARAL finds each region in time proportional to a single alignment.

Parametric Alignment

Ralf Zimmer, GMD, Germany

Parametric alignment is a systematic approach to determine optimal parameter values for a given optimization procedure (called an oracle) and a set of parameters.
First, we determine a tessellation of the parameter space into regions where the same solutions are optimal. This can be done easily for quite arbitrary optimization problems as long as the objective function depends linearly on the parameters to be estimated. In this case any solution $a \in A$ defines a linear function h^a . The problem is to computer the upper envelope $\max_{a \in A} h^a$ and its projection onto the parameter space of these functions. The latter decomposes the space into convex polytopes. Our standard application uses a sequence alignment oracle, but we are also interested in optimizing parameters for structure comparison and sequence–structure alignment (threading) algorithms. We present several algorithms for different types of oracles, which deal with or try to avoid numerical problems

(the oracle can provide optimal solution only for certain parameter combinations and with limited accuracy), minimizes the number of oracle calls and generalizes to arbitrary dimension, i.e. arbitrary number of parameters.
Second, having all possible optimal solutions for any combination of estimated parameters we compare against a set of correct solutions to derive parameter values which best reproduce them. For assessing the performance of a particular optimization method we introduce several distances to measure deviations of any of the optimal solutions from the correct solution. The analysis revealed that for standard method and scoring systems there is still a large discrepancy between what could in principle be achieved by such a method and what is actually achieved with fixed — even optimized — parameter values.

Using Simple Markov Models to Search DNA Databases

Kevin Karplus

The talk presented a technique for using simple Markov models to search for interesting sequences in a database of DNA sequences.
Two techniques (neighborhood and complement blurring) are introduced to help the model generalize from small training sets.
Model was used for finding REPS (repetitive extra-genic palindromic sequences) in E.coli. Results were about as good as for known techniques (& much faster).
Models can also be used to provide a starting point for building hidden Markov models (explained in tech-report UCSC-CRL–94–24 available by world–wide–web, URL http://www.cse.ucsc.edu/~karplus).

Reconsidering Evolutionary Tree Construction

Tandy Warnow, University of Pennsylvania, USA

We consider the problem of evolutionary tree construction and posit that the standard techniques based upon optimization criteria when applied to aligned biomolecular sequences will produce larger solution sets (i.e. too many candidate trees). This means we will need to develop better ways of representing and resolving ambiguity. We present two new consensus methods. The *local consensus* (SODA '95) and the *asymmetric median tree*. The *asymmetric median tree* of a profile $\{T_1, T_2, \ldots, T_k\}$ is a tree minimizing

$$\sum_{i=1}^{k} |C(T_i) - C(C_T)|$$

where $C(T_i)$ denotes the binary encoding of the tree T_i (i.e. set of binary characters given by the edges in T_i). We show that any *asymmetric median tree* is at least as informative as any median tree or the strict consensus and equals the compatibility tree ($C(T) = \bigcup C(T_i)$) when it exists. We also discuss the perfect phylogeny problem (PP) and show that the algorithm of Kannan & Warnow for PP can be used to solve evolutionary tree construction then almost all characters are compatible. We cite an application of this algorithm to Indo-European data, in which all but 5 out of 59 informative characters are compatible. We then suggest that data comparable to this linguistic data could help solving biological evolutionary tree problems.

Perspectives on the Statistics of Local Alignment Score

Stephen Altschul, NCBI, NLM, NIH, USA

Given a particular scoring system, a central question has always been how high a score may be expected to occur purely by chance. Computational experiments suggest that the optimal scores of gaped sub-alignments follow an extreme value distribution. We consider here the dependence of the *characteristic value* n of the distribution on the comparison size N (the product of the length of the sequences compared).
For some measures of sequence similarity, n is known to grow asymptotically as $\log N + C$, while for others n grows as $\log N + A \log\log N + C$, where the base of the logarithm is $e^{-\lambda}$, and λ is the decay constant for the extreme value distribution. If n is plotted against $\ln N$, the presence of a $\log\log$ term would result in a fitted line with slope greater than $1/\lambda$. For the scores of sub-alignments allowing gaps, this is indeed what is observed. However, an edge-effect correction, based upon the observed lengths of optimal sub-alignments, makes the slope discrepancy disappear. This supports a conjecture of Waterman & Vingron that n grows as $\log N + C$, as has been established analytically by Demko & Karlin for sub-alignments without gaps. It allows a computational experiment for a single N to estimate the relevant statistical parameters for all N.
This is joint work with Dr. Warren Gish.

Tutorial: Protein Structure

Chris Sander, EMBL Heidelberg, Germany

Some facts about proteins:

1. Different positions are subject to different evolutionary selective pressures.
   (a) Active sites
   (b) Support near active sites
   (c) Structural scaffold (interior core)
   (d) Loops evolve at $\tau_a > \tau_b > \tau_c > \tau_d$ (τ = time interval)
2. Protein folding is cooperative.
3. There are many–to–one relationships $sequence \longrightarrow structure$ and $sequence \longrightarrow function$.
4. Protein folding is ... (exercise left to the reader)

Challenge: Develop a quantitative theory of protein evolution at the residue, protein, cell and organism level!

Visualization Procedures Related to Abstract Similarity Analysis, Quasi–Crystals, and the Traveling Sales-man's Problem

Andreas Dress, Universität Bielefeld, Germany

Based on the SPLIT-DECOMPOSITION technique, visualization procedures have been developed, which visualize abstract similarity relationships in terms of the family of associated weighted splits. They have been applied successfully to biological as well as psychological and other data. They rely on *appropriately chosen* projections of *appropriately chosen* subgraphs of the N-dimensional hypercube representing simultaneously all splits in the family (N being the number of these splits). Consequently, the visualization procedures are closely related to deBruijn's grid dualization method for generating Penrose patterns and other quasi–crystalline structures. Using this relationship, one can find necessary and sufficient for the split system related to an abstract dissimilarity matrix to have a generically (that is, weight independent) planar representation. Amazingly enough, these conditions are also necessary and sufficient for a certain lower bound which can be defined for every TSP instance to be actually achieved by the TSP associated with the given dissimilarity matrix, provided the triangular inequality is satisfied, — a relationship which is crucial for the applicability of deBruijn's grid dualization method in our context as this requires, as an input,

an appropriate cyclic ordering of our objects which turns out to be precisely that cyclic ordering which minimizes the length of the traveling sales-man's tour.

Applications of Scoring Functions in Sequence Comparison and Evolutionary Studies

Volker Brendel, Stanford CA, USA

For a sequence of scores $X_1, X_2, \ldots, X_N$ under broad assumptions on the underlying distribution of the X_i ($Prob\{X_i = S_k\} = p_k$ for all $i; max\ p_k > 0; \sum p_k s_k < 0$; and generalizations) statistical theory defines the threshold S_p of the maximal aggregate score of a sequence segment to be significant at the p-significance level. The theory is being used to establish benchmarks for detecting regions of unusual composition within proteins as well as segment pairs of high similarity comparing two proteins (HSSPs – high scoring segment pairs). Substitution scoring matrices can be defined that give characteristic expected features of HSSP, (percent identity, length). Sequence comparisons can be based on optimally consistently ordered HSSPs, with regions between the HSSPs left unaligned and un-scored (SSPA method – significant segment pair alignment). Pre-calculated HSSPs can be used to give positive weights to conserved residue pairs in the HSSPs in a complete (dynamic programming) alignment allowing gaps. This procedure (SSPALI) essentially faces the (global) alignment to go through the HSSPs and eliminates sensitivity to gap penalties.

Amino Acid Pair Interchanges at Spatially Conserved Location

Dalit Naor, Tel Aviv University, Israel

The pattern of Amino Acid pair interchanges which occur at spatially, locally conserved regions in globally dissimilar and unrelated proteins is studied. The spatially conserved, structural motifs consist of a *large enough* number of C_α's found to provide a geometric match between 2 proteins, regardless of the order of the C_α's in the sequence, or of the sequence composition of the substructures. The motifs were obtained by applying a computer–vision algorithm, the Geometric–Hashing.
The interchanges at geometrically similar positions demonstrate the expected behavior. Yet, a closer inspection reveals some distinct characteristics, as compared with interchanges based upon sequence–order technique or from energy–contact–based considerations. These differences can be explained in terms of

the 3–dimensional structures of the proteins. Most of all, there is a clear distinction between residues preferring to be on the protein surface, compared to those frequently buried in the interior. Our data are presented in the form of a 20×20 amino–acid substitution matrix. The relative entropy of this matrix is low, making it unfit for most homology searches. However, this structure–based amino–acid sequence–order–independent matrix has a range of applications to protein structure prediction methods, such as threading and modeling.

Clustering of Protein Structures on the Base of Hexapeptides

Joachim Selbig, GMD, Germany

Protein structure prediction by threading relies mainly on the reduction of the structure description dimensionality. Within the context of 2D structure representations different contact descriptions are used varying in their complexity from about 100 to about 100 000. There is the question whether the chosen description parameters (sequential distances, spatial distances, etc.) meet such observations that structurally similar proteins can have as few as 12 % of their interactions in common. An answer to this question is given, of course, by the performance of the fold recognition system. But it would be of interest to evaluate the reduced structure description in advance.
We consider contact patterns as descriptions of characteristic interaction patterns which, in particular, are formed by contacts between residues that are far apart in the sequence. In order to evaluate the contact patterns in advance we apply them to the global comparison of protein 3D structures. N proteins characterized by M contact patterns define an $N \times M$ data matrix X whose entries x_{ij} indicate the number of instances of contact pattern j in protein i. The similarity of two proteins k and l is defined by the Euclidean distance between their vector descriptions $x_k = (x_{k1}, \ldots, x_{kM})$ and $x_l = (x_{l1}, \ldots, x_{lM})$. This distance allows the clustering of the given set of proteins. Different standardization techniques of the raw matrix X and different measures of distances between clusters are discussed.

Detecting Functional Regions in Protein Sequences

Georg Casari, EMBL Heidelberg, Germany

During the times of evolution proteins are subjected to random mutations and selections. The traces of this evolutionary process can be observed by comparing several members of a protein family with a common ancestor. At positions

of functional importance the selective pressure is very high and generally these residues are conserved throughout the protein family of sequences. More subtile patterns of conservation arise at positions, where some members have acquired new specificities and new biological roles. Although not immediately obvious from the comparison of sequences, these *specificity determining* residues are crucial.
We present a new representation of protein families. The basis of this new representation is a multiple alignment of the protein sequences that has been generated by any standard program or by careful manual intervention. This alignment is recoded in a set of vectors in *Sequence Space*. Those vectors are highly dimensional (20× length of alignment). By principal component analysis the most informative dimensions for the particular sequence family are calculated and low dimensional representations of proteins as points in Sequence Space are obtained. These representations maximally preserve similarity relationships between the proteins as spatial proximities. By a simple projection it is possible to obtain the complete sequence information in the same set of principal axes. This view highlights conserved residues as those taking extreme positions. Different directions in this sequence space correspond to different specificities and associated sequence signatures can be identified and predicted to contain the functional residues. The method was illustrated by the superfamily of GTP–binding domains of Ras-like proteins, analyzed with sequence space and compared to experimental facts.

Computing Protein Cores Using Probabilistic Representations

Russ Altman, Stanford University, USA

We have developed a method for computing the spatially invariant (or low variance) core from a multiple alignment. Our cores are a proper subset of all alignable positions and are characterize by a homogeneous distribution of the volumes of spatial variation. The volumes are the 3D Gaussian distributions that enclose the position of equivalent atoms.
We find a globin core with 12 atoms that has numerous biological correlates in folding path, other protein families and exon organization. We find an immunoglobin core with similar biological correlates. We are in the process of computing cores of automatically aligned protein families of Sali and Overington and the FSSP database of Holm and Sander. We anticipate this will be available in fall 1995.

Protein Structure Database Searches

Liisa Holm, EMBL, Germany

With a rapidly growing pool of known tertiary structures, the importance of protein structure comparison parallels that of sequence alignment. The Dali algorithm is based on comparing $C_\alpha - C_\alpha$ distance matrices. The distance matrices are find decomposed into elementary contact patterns, e.g., hexapeptide–hexapeptide submatrices. Then, similar contact patterns in the two matrices are paired and combined into larger sets of pairs. A Monte Carlo procedure is used to optimize a similarity score defined in terms of equivalent intramolecular distances. Several alignments can be optimized in parallel, leading to simultaneous detection of the best and suboptimal solutions. The method allows sequence gaps of any length, reversal of chain direction, and free topological connectivity of aligned segments. Sequential connectivity can be imposed as an option. The method is fully automatic and identifies structural resemblances and common structural cores accurately and sensitively, even in the case of geometrical distortions. All–against–all alignment of representative protein structures results in an objective classification of known 3D folds. Unexpected topological similarities of biological interest have been detected.

Finding the Most Surprising Structural Similarities

Steve Bryant, NCBI, NIH, Bethesda, USA

We present a fast algorithm and significance statistic for structure–structure similarity search. Protein structures are represented by there *core* alpha helices and beta strands, each described by an vector. A substructure is defined to be similar, between two proteins, when these vectors may be superimposed well. We identify candidate substructure similarities by constructing a comparison graph in which nodes represent aligned vector pairs and edges the presence of a sensitive spatial and chain–order relationship, in the manner of Artymuik and colleagues. Cliques are then identified and ranked according to the statistic $L = \prod_i p_i$, where p_i is the probability that the RMS residual observed for the corresponding secondary structure element pair would be observed by chance, as judged by the appropriate empirical distribution. Statistical significance is given by the expected number of cliques with this score, $E = DNL \times p$, where N is the number of like–size vector combinations possible, given the sizes of the two proteins, and the number of proteins in the database searched. We show that this algorithm identifies known structural similarities with 99 % sensitivity and with search times of approximately 1 minute for a 2000–protein data base. Residue alignments are refined by

Monte Carlo, and appear to agree perfectly with iterative values.

Threading by Recursive Dynamic Programming

Ralf Thiele, GMD, Germnay

We propose a new alignment procedure that is capable of mapping a sequence into a given structure optimizing some interaction potential. Recursive Dynamic Programming (RDP) is a hierarchical method which, on each level of hierarchy, identifies locally optimal solutions and assembles them into partial alignments between sequences and structures.
In contrast to classical dynamic programming, RDP is designed to handle objective functions not obeying the principle of prefix optimality, e.g. scoring schemes derived from energy potentials of mean force. For such alignment problems RDP aims at computing alignments that are near – optimal w.r.t. the involved cost functions and biological meaningful at the same time. Towards this goal RDP maintains a dynamic balance between different factors governing alignment fitness such as evolutionary relationships and structural preferences.
As in the RDP method gaps are not scored explicitly, the problematic assignment of gap cost parameters compatible with interaction potentials is circumvented. In contrast to fragment threading approaches RDP addresses the full threading problem. The important parts of proteins, e. g. active sites, often are located in loop regions and are highly conserved if two proteins share function and structure. This evolutionary conservation guides the RDP procedure, but is ignored by approaches only mapping secondary structure elements or core regions. In order to evaluate the RDP approach we analyzed whether known and accepted multiple alignments based on structural information can be reproduced with the RDP method.

An Alignment Model and Fast Algorithm for Protein Threading

Steve Bryant, NCBI, NIH, Bethesda, USA

We propose an explicit definition of the *core* of a protein structure as a series of chain–continuous segments, each corresponding to an alpha helix or beta strand. Threading of a sequence through this structure, for purposes of fold recognition, may then be represented as alignment of each core element with an un-gaped block of the sequence. With only one alignment variable per core element, we find that the space of alternative alignments may be efficiently searched by a Monte Carlo procedure. Furthermore, we find that the optimal boundaries of

core elements may be easily identified by introduction of two additional *extent* variables per core element. These algorithms have been tested in control threading experiments involving known structures and sequences. Significant threading scores were obtained when roughly 60 % of residue sites were conserved between two proteins, regardless of the degree of sequence similarity. Accurate alignments are obtained when RMS superposition residuals are under 3 Å. This algorithm behavior was also seen in a series of blind predictions undertaken for the recent Asilomar workshop. Accurate alignments were predicted when RMS residuals were under 3 Å, and more than 50 % of sites superimposable.

Teaching Computational Molecular Biology at Stanford

Russ Altman, Stanford University, USA

We have just finished teaching a course entitled "Representations and Algorithms for Molecular Biology" at Stanford. The course is geared towards students in Medical Information Science and Computer Science. It is a programming course with 7 small ($1-3$ hour) assignments and 3 larger projects (>20 hours), including dynamic programming sequence alignment, RMS fitting of 3D structure and threading to identify globins vs. non globins.
The course was very well received and is now mandatory for Medical Information Science graduate students.

Distance Education on Biocomputing via the Internet

Robert Giegerich, University of Bielefeld, Germany

A group of 7 instructors and 35 students from many countries participate in a BioComputing course held on the Internet. The course is based on a textbook in hypertext, directly cross-linked to numerous biocomputing resources on the WWW. Weekly class sessions in small groups are held in the virtual laboratory BioMOO, which provides many–to–many communication tools and a tutorial pairwise alignment machine. For more information on the course, see

http://www.techfak.uni-bielefeld.de/bcd/

to learn more about BioMOO, use

telnet bioinformatics.weizmann.ac.il 8888.

Generating New Phylogenetic Information from a Collection of Trees

Mike Steel, Mathematics Department U. of Canterburg, New Zealand

Suppose we have a collection of (phylogenetic trees) whose leaves are labeled from overlapping sets of species. Such a collection may or may not be realized by a single *parent* tree (and deciding this is in general NP-hard), but when these trees do fit together coherently, they often force other subsets of the species to form a particular phylogenetic tree (not present in the input).
By formalizing how these new relationships are derived, we can address a number of natural questions and, in particular, settle two conjectures raised by M.C.H. Dekker, concerning the hierarchy of *rules* one can obtain for generating new trees. The approach exploits the property that, when all of the input trees share a leaf, then the properties we are interested in translate into properties of an associated graph.
In the last part of the talk I also describe some recent work that allows a phylogenetic tree to be recovered from (sufficiently long) sequences that evolve $ii\delta$ under a minimal Markov assumption (and no restrictions on the underlying transitions matrices M_e, beyond $\det M_e \neq 0, I1$).

Combinatorial Problems in Physical Mapping

Ron Shamir, Tel Aviv University, Israel

Constructing physical maps is essential to many genomic challenges and in particular to the Human Genome project. Given a set of intervals and their pairwise interactions, the goal is to reconstruct their relative order or determine that none exists. With perfect data, the problem is linear. With most models allowing errors, the problem becomes NP-hard. However, by introducing additional biological constraints, the problem becomes polynomial.
We also discuss problems of realizing interval graphs when additional size, order or distance constraints are present. Some problems are surprisingly hard while for others we have linear time algorithms.

An Evolutionary Mixture for Describing Protein Sequence Families

Michael Gribskov & Stella Veretnik, Supercomputer Center, San Diego, USA

There are a number of interesting problems related to families of proteins including:

classification: Is this sequence a member of a known family?

alignment: How does one map a sequence onto a family residue–by–residue?

inference: How does one describe the important general features of the family based on a small biased sample ?

structure vs. evolution: How can one separate these effects in the pattern of conserved residues seen in a family ?

In this work we describe an approach to describing protein families with a position–specific scoring system (or profile). The evolutionary profile is an attempt to overcome the defects of the Dayhoff evolutionary model and to give a biologically relevant description of a sequence family.
The evolutionary profile is a finite mixture model based on component distributions drawn from the Dayhoff evolutionary model (PAM model). Each aligned position in a group of sequences is compared to model distributions for the 20 possible ancestral residues at evolutionary distances $2^n, n = 0, \ldots, 10$, PAM, (220 model distributions in all). For each ancestral residue the best matching distance is selected as the one that maximizes the relative entropy, H, of the model and observed distributions $H = \sum_{i=1}^{20} f_i \; ln\frac{f_i}{p_i}$ where f_i are observed frequencies of residue i and p_i are the model frequencies from the Dayhoff model.
This reduces the 220 model distributions to 20, one for each possible ancestral residue.
These 20 model distributions are combined into a mixture model with weights related to the probability that the model for ancestor a, M_a, explains the observed data, F, and that the background probability of a random model M_r explains the observed data:

$$W_a = [P(M_a|F) - P(M_r|F)] \; / \; \sum(P(M_a|F) - P(M_r|F)) \text{ and}$$

$$P(M_a|F) = P(\text{Ancestor} = a|F) = \frac{P(\text{Ancestor} = a)(P(F|M_a)}{\sum P(\text{Ancestor} = a)(P(F|M_a)}$$

note that the random distribution, M_r, is simply the expected residue distribution and is equivalent to a Dayhoff distribution at distance ∞. The weights, W_a, are therefore always positive since $(P(M_a|F) \geq (P(M_r|F)$.

The evolutionary profile is defined as the log-odds form of the weighted mixture:

$$Profile(i,j) = ln\left[\frac{\sum_{a=1}^{20} W_{ai} P_{aij}}{P_{rj}}\right]$$

The accuracy of evolutionary profiles as protein family classifiers have been evaluated by cross–validation studies using Receiver Operating Characteristic (ROC) analysis of local similarity dynamic programming (Smith–Waterman) database searches. Studies on the $4Fe-4S$ ferredoxin and ATP dependent RNA helicase families show that the evolutionary profile method is significantly better than the earlier average profile method (Gribskov *et al.*, 1987), and has less than half the misclassification error. More surprisingly, we find that high discrimination can be achieved with profiles generated from as few as two or three sequences. This strong ability to generalize from small datasets is due to the incorporation of biologically relevant prior information from the Dayhoff model.

An Algorithm for Finding All Maximal Common Substructures and Its Application to Protein Structures

Ina Koch, GMD - SCAI, Sankt Augustin, Germany

For the comparison and analysis of protein structures, it is of interest to find maximal common substructures in a given set of proteins. This question is relevant for a suitable well–formed definition of the topology of secondary structure motifs in proteins, for instance.
We describe a suitable representation of the secondary structure topology of a protein by an undirected labeled graph. Then we transformed the maximal common subgraph problem in two graphs to the maximum clique problem in one graph. We develop an algorithm which bases on the method by Bron & Kerbosch which enumerates all maximal cliques in a graph. The main improvement of our algorithm is to restrict the search process to cliques which represent connected substructures. This restriction reduces the size of the search tree drastically. While the graph problem we have to solve is NP-hard the algorithm can handle graphs that arise from actual biological data within a matter of a few seconds.

On the Prediction of RNA Secondary Structure

Norbert Blum, University Bonn, Germany

Algorithms for the prediction of RNA secondary structures were presented. The considered energy rules are context–dependent. In the case that the energy rules are concave, we have obtained algorithms which use for bulges $O(n^2)$ and for interior loops $O(n^3)$ time, where n is the length of the primary structure of the RNA molecule under consideration.

Constraint Satisfaction Methods Applied to the Solution of the 3D Structure of the 30 S Ribosomal Subunit

Russ Altman, Stanford University, USA

The problem of finding the structure of the 30 S ribosomal subunit in prokaryotes can be stated as the task of positioning approximately 60 oriented cylinders in a framework of 21 globulin proteins (whose position is known from neutron diffraction). We generate a representative sample of the set of conformations of the cylinders (each of which represents an A–form RNA double helix). The system works hierarchically by generating possible positions for helices, pruning this lot of positions using constraints between helices, and then combining partial solutions to create *coherent instances* of conformations that satisfy all experimental and chemical constraints. We use a constraint satisfaction formulation of the problem and maintain tractability by

1. performing constraint satisfaction on groups of nodes with low count before those with high count;
2. performing intelligent sampling of conformation to maintain the *most different* conformations;
3. varying the precision of sampling to keep location counts low.

We show that our method allows us to generate multiple alternative conformations for the 30 S subunit. They share certain critical features, including a notch into which we have docked tRNA preliminarily. We are building a knowledge base of ribosomal structural data to support our model building efforts. This work was done in collaboration with Harry Noller at U.C. Santa Cruz.

Representing Large RNA Folding Landscapes

Robert Giegerich, University of Bielefeld, Germany

In order to validate a mathematical model of evolution along paths of neutral mutations, the sequence–structure map for a sequence space of size 2^{30} was computed in its entirely. The talk explains the motivation behind such a computational experiment and describes the data structures designed for connectivity analysis in very large graphs of neutral nets.

Computer Models of Ultraselfish Genes in Mice

Dannie Durand, Bellcore Morristown, NY, USA

The *t-haplotype* is an ultraselfish region on chromosome 17 in mice that propagates itself at the expense of its allelic partners through unusual sperm killing properties. We wish to study how such a system can evolve and persist in nature. Because of the complex social structure of mice, gene frequencies cannot be modeled analytically and computer simulation must be used.
We built a simulation of the population genetics of this non–Mendelian system based on our current understanding of its biology. By comparing predicted gene–frequencies with those measured in wild populations, we can evaluate the biological model and propose new biological experiments to refine it.
Thus computer simulation of population genetics is a tool which allows us to study a system that has a molecular biology component, a behavioural component and a genetics component. The design of this simulator raises interesting computational problems requiring combinatorial solutions.
This work is in collaboration with Professor Lee Silver of the Department of Molecular Biology at Princeton University.

Parallel Protein Puzzle

Hans-Peter Lenhof, MPI für Informatik, Germany

We have implemented a parallel distributed algorithm for the geometric docking problem which uses a new measure for the size of the contact area of two molecules. The measure is a potential function that counts the *van der Waals* contacts between the atoms of the two molecules. The algorithm almost always found good (RMS) approximations of the real conformations that were below the best five dockings. In 42 of 52 test examples the best conformation with respect to the potential function was an approximation of the real conformation. The running time of our sequential algorithm is in the order of the running time of the

algorithm of Neul *et al.* The parallel version of the algorithm has a reasonable speedup and modul communication requirements.

Time–Efficient Docking of Flexible Ligands

Matthias Rarey, GMD - SCAI, Sankt Augustin, Germany

Most biochemical processes in living systems are based on the specific binding of small organic molecules (ligands) to the active site of proteins (receptors). A major goal of pharmaceutical research is to control such processes by designing molecules with high binding affinity to a given receptor molecule. If the three–dimensional structure of the receptor is known, rational drug–design techniques are applicable. The docking problem is a key–problem in rational drug design:

> Given a three–dimensional structure of a receptor and the structure of a ligand, predict the binding affinity of the ligand to the receptor and the geometry of the receptor–ligand complex.

In this talk, we describe an algorithm for placing flexible ligands in active sites of proteins. The two major goals in the development of our docking method, called FlexX, are the explicit exploitation of molecular flexibility of the ligand and the development of a model of the docking process that includes the physico–chemical properties of the molecules.
Our docking method consists of three phases: The selection of a base fragment, the placement of the base fragment in the active site and the incremental construction of the ligand inside the active site. Except for the base selection, the algorithm runs without manual intervention.
For an increasing set of test cases, the algorithm is able to reproduce complexes known from X–ray crystallography in a few minutes on a workstation with an error bound of about 1.5 Å RMS or less.

A Database System Supporting Protein Docking – A 3D Molecular Surface Representation

Thomas Seidl, University of Munich, CS Institute, Germany

Protein–Protein Docking is a new and quite challenging application for spatial database systems, even with the focus of $1-n$ docking problem: given a query protein A and a protein database, e.g. PDB, docking partners from the PDB interacting with A should be returned. Thus, 3D structures and surfaces of molecules become basic objects in molecular databases. As a well–known technique in spatial database systems as well as in molecular biocomputing, the multi–step

query processing paradigm seems to be quite appropriate. We determine surface features to representative geometric and physico–chemical properties of surface points and their neighborhood. Therefore, neighborhood determination has to be provided as a basic operation, along with an appropriate molecular surface representation. We suggest a patch–based data structure, called the *tri–edge* structure, first, to efficiently support neighborhood query processing, and second to save space in comparison to common 2D subdivision data structures such as the *quad–edge* structure or the *doubly–connected* edge list. This way, we provide efficient access to the well known Connolly subdivision of molecular surfaces, in particular we support topological access such as traversal of the patches.

Computing Optimal Multiple–Sequence Alignments

John Kececioglu, University of Georgia, USA

We study the exact solution of a new problem in multiple–sequence alignment, which we call *maximum-trace alignment*. Informally, the input is a set of pairs of matched characters from the sequences; each pair has an associated weight. The output is a subset of the pairs of maximum total weight that satisfies the following property: there is a multiple alignment that places each pair of characters selected by the subset together in the same column. A set of paires with this property is called a trace. Intuitively, a trace of maximum weight specifies a multiple alignment that agrees as much as possible with the character matches of the input.
We develop a branch-and-bound algorithm for maximum-trace alignment that exploits the combinatorial structure of the problem, and proceeds by solving a series of minimum-cut-problems. Though maximum-trace is NP-complete, our experience with an evolving implementation shows we can at present solve to provable optimality instances on 5 very distant sequences of lenght 200 (for example, from M. McClure's retrovirus benchmark dataset). These are the largest instances that have been solved to optimality to date for any formulation of multiple alignment.

Rapid Similarity Search

Gene Myers, University of Arizona, USA

We begin with a detailed development of the search algorithm in BLAST and some of the history of its development. We emphasize the concept of *neighborhoods* as a way of increasing selectivity without compromising sensitivity. While

BLAST is heuristic with respected to its stated computional criterion, this author simultaneously was using the neighborhood idea to develop a deterministic and provably sublinear, off–line algorithm for approximate keyword matching. The algorithm was sketched and its connection to BLAST exposed.

Bonsai Garden: Parallel Knowledge Discovery Systems for Sequences

Satoru Miyano, Kyushu University, Japan

We have developed a machine discovery system BONSAI which receives positive and negative examples as inputs and produces, as a hypothesis, a pair of a decision tree over regular patterns and an alphabet indexing. This system has succeeded in discovering reasonable knowledge on transmembrane domain sequences and signal peptides sequences by computer experiments. However, when several kinds of sequences are mixed in the data, it does not seem reasonable for a single BONSAI System to find a hypothesis of a reasonable small size with high accuracy. For this purpose, we have designed a system BONSAI Garden, in which several BONSAI's and a program called *Gardener* run over a network in parallel, to classify the data into some number of classes and, simultaneously, to discover knowledge for each of these classes. Computional experiments in BONSAI Garden show an interesting observation.

SRS – A Retrieval System for Molecular Biological Databanks

Thure Etzold, EMBL Heidelberg, Germany

We develop a retrieval system (SRS – Sequence Retrieval System) for databanks in a flat file format. Before query time the input databanks are read and parsed and words extracted that are put into indices, usually one for each data–field. The information SRS needs for processing a databank, such as the file structure, the data-fields and their syntax is specified in a single file using two languages, ODD (Object Design and Definition) and ICARUS (Interpreter for Commands And Recursive Syntax).
The crossreferences that most biological databanks provide are also processed before query time to build link indices between pairs of databanks which, when combined, give rise to an entire network where the links are the edges and the databanks the nodes. In this network it is possible to navigate from any databank to any other either by a single or a sussesion of links.

A WWW server for SRS has been developed and is currently on $\sim$ 15 nodes serving a total of $\sim$ 80 databanks. One of them is the server at the EMBL Heidelberg with the URL http://www.embl-heidelberg.de/srs/srsc.

Loglinear Models for Splice Site Recognition

Jürgen Kleffe, FU Berlin, Germany

Prediction of splice sites is an important part of all algorithms for exon–intron structure recognition. Given a sequence, potential introns are identified, removed and the remaining sequence is matched against the information available for coding sequences in general or in particular, depending on the problem. Commonly, the algorithm works for optimal matches and the number of exon–intron structures to be considered grows quickly with the number of sequence fragments that are considered as potential introns. It is therefore important to have splice site prediction methods which make as less as possible false positive predictions while missing non of the true sites. Each true site missed causes the exon–intron structure prediction to fail and therefore splice site prediction methods used in such algorithms are best tuned to make no false negative predictions. We discuss data where under this condition 25 % less donor sites are wrongly predicted based on loglinear and logit-linear models as compared to methods based on weight matrices and likelihood ratios over splice site positions. The gain comes from a better use of interactions between nucleotides in distant splice site positions.

Sequence Alignment and Phylogeny

Martin Vingron, GMD St. Augustin, Germany

Multiple sequence alignment programs usually calculate a rough phylogenetic tree to guide an iterative alignment procedure. The resulting alignment, however is in turn used to deduce phylogeny. In an attempt to avoid this circularity we develop a heuristic approach to integrate tree construction and alignment. The suggested procedure mimics known heuristics for the approximation of additive metrics in that one sequence is added at a time. Simultaneously, the alignment is updated such that both tasks (tree construction and alignment) are performed. First results are encouraging and a fast, practical version of the program is under development.

How to Think about Repeats in Sequences

Dan Gusfield, University of California, Davis, USA

Repeated substrings are ubiquitous in DNA, and even in proteins one wishes to analyse related proteins that arise by duplication and modification. However, unless one is careful with the definitions of a repeat, it is easy to be overwhelmed by *flat* information about repeated substrings. In this talk I examine how one can organize information about repeats using a suffix tree. I show how to efficiently compute a partial ordering *Maximal repeats*, how to find all the *Super–maximal repeats*, how to compute the degree of *near supermaximality*, and how these notions might be used to define distances and signatures for biological sequences.