A Paradigm Shift in Biological Computing
- A Survey on the Development of Bioinformatics Algorithms in Taiwan

Wen-Lian Hsu
Institute of Information Science, Academia Sinica, Taipei, Taiwan, R.O.C.

In this survey we focus on bioinformatics literature involving computational algorithms from authors in Taiwan.

In a typical scenario of discrete algorithm design we have a single objective to optimize or a certain property to recognize. One can expect either to give a polynomial time algorithm for the problem or to show that it is NP-complete/hard. However, biological experiments are never flawless. The data involved in biocomputing algorithms is always noisy. Thus, the traditional rigid discrete algorithm does not seem to produce a viable solution for biologists.

For an optimization problem it is sometimes possible to ask how close one can approximate to an optimal solution. However, it is more troublesome for a recognition problem given the noisy data. One can certainly modify the noisy version to an optimization problem by asking what is the minimum number of modifications (e.g. deletions, insertions or substitutions) to transform the noisy data into one that conforms to the prescribed property. But such a purely combinatorial approach usually suffers from the following two unpleasant phenomena: 1. The problem of finding this least number would likely become *NP*-hard; 2. Even if one can find the best modification in this sense, the data in the resultant configuration might not make any biological sense.

There are many other complications to a biological computation problem such as multiple objective optimization with dependent objective functions, heterogeneous constraint types such as subset selection constraints, numerical weighting constraints, plus local noises that are less amenable to a global optimization approach. To resolve this, the key is to employ as much problem structures (both local and global) and features as possible in the algorithm design. Therefore, data dependent algorithms such as natural language and machine learning techniques are becoming increasingly more important. We will make an overview in the following areas: sequence analysis, sequence/structure alignment, protein structure prediction, phylogenetic analysis, motif finding, gene finding/annotation, protein structure prediction, protein structure identification and literature mining.

Sequence analysis

Physical mapping has been a long-standing difficult problem troubling biologists until

the shotgun sequencing approach becomes popular. The mathematical problem underlying physical mapping is the consecutive ones test. Although Booth and Lueker's PQ-tree can solve this problem, the technique falls apart in case there are errors. Conventional approach almost always assumes that certain errors do not occur, which renders them impractical for real data. Dr. Hsu's group has modified previous rigid algorithms for testing consecutive ones property into one that can accommodate clustering techniques, and produces satisfactory approximate probe orderings for most data [Lu et al 2004]. Similar approach has also been applied to test interval graphs under noise [Lu et al 2003b]. Their results can be used for real world data for physical mapping and clone assembly, two major problems in DNA sequencing.

Dr. Hwang's group develops a uni-marker approach to locate SNPs in a genome-wide sequence [Chen et al 2002]. A uni-marker is a DNA fragment of length 15 that appears only once in a whole genomic DNA sequence. By using this approach, one can locate SNPs efficiently without using any BLAST search. Dr. Hsu's group extends the notion of uni-marker to align ESTs against a whole genome sequence [Hsu & Chen 2003].

Repeat sequences are usually associated with biological meanings. Dr. Hrong's group has creates a repeat sequence database, and applying data mining techniques to find useful elements, such as regulatory element in promoter regions [Horng & Cho 2000; Hrong & Huang 2002; Hrong et al 2002]. To mine these data, they use many probabilistic/statistical approaches to generate rules that are used to distinguish special regions from others.

To find specific regions of biological sequences (e.g. GC rich regions), Dr. Chao's group develops some efficient algorithms to find *right-skew decompositions* of strings. For a given string, if the average of any prefix is always less than or equal to the average of the remaining suffix, we say that the string is *right-skew*. The definition of right-skew decomposition is as follows. For a given string S, partition S into substrings S_1, S_2,…,S_k such that each S_i is a right-skew substring of S, and we have density(S_1) > density(S_2) > … > density(S_k). With the aid of right-skew decomposition, many biological sequence problems can be solved efficiently and effectively [Lin et al 2002, 2003; Lu et al 2003a].

We also define a new similarity measure of protein sequences that is based on small peptide fragment comparison rather than amino acid comparison; since we believe that biological meaning is associated more with small peptide fragments than individual amino acids [Wu et al 2003].

Sequence/Structure Alignment

For a general survey, readers can consult [Lassmann & Sonnhammer 2002; Notre-

dame 2002; Thompson et al 1999].

Dr. Chao's group has done extensive works in sequence alignment. To align a cDNA against a genome DNA sequence, they use several string-matching techniques to compute an alignment with restricted affine gap penalties efficiently [Chao 1999]. Since homologous sequences are sometimes similar over some regions but different over other regions, they define a generalized global alignment model to handle sequences with intermittent similarities, and design a dynamic programming algorithm to compute an optimal general alignment [Huang 2003].

Most alignment algorithms cannot handle large scale sequence due to computation power. Dr. Shih and Dr. Li have proposed an alignment algorithm that can align whole genomic sequences [Shih & Li 2003]. They first encode whole genomic sequences into coded sequences, and then perform efficient comparison/alignment procedure on the coded sequences. Their algorithm can also handle repeat regions.

Dr. Tang's group design a method for computing a constrained multiple sequence alignment for guaranteeing that the generated alignment satisfies the user specified constraints that some particular residues should be aligned together [Tang et al 2003; Tsai et al 2004]. They design a constrained pairwise sequence alignment algorithm, then create a minimum spanning tree of the sequences, and align the sequences progressively (using the constrained pairwise alignment algorithm) according to the tree.

Protein structure comparison has been used widely in the study of structural and functional genomics. However, it is computationally expensive and as a result almost all of the methods currently in use only look for the optimal alignment and ignore many alternative alignments that are statistically significant and that may provide insight into protein evolution or folding. Dr. Hwang's group has developed a new protein structure comparison method to detect potentially viable alternative alignments in all-against-all database comparisons [Shih & Hwang 2003]. They rank alignment solutions based on derived secondary structure element-matching probabilities.

Phylogenetic Analysis

For a general survey, readers can consult [Kim & Warnow 1999; Kao 1998].

There are two main research topics in phylogenetic analysis: phylogenetic tree construction and phylogenetic tree comparison. Dr. Lin and Dr. Hsu propose an algorithm to find isomorphic subtrees of phylogenetic trees to exploit evolutionary information [Lin & Hsu 2003].

Dr. Tang's group develops many tree-related algorithms such as spanning trees and steiner trees [Lu et al 2000; Lu et al 2003c; Wu et al 1999]. These algorithms can be used to perform phylogenetic analysis.

Motif Finding

For a general survey, readers can visit the following web site:
http://www.biotech.ufl.edu/WorkshopsCourses/bioinfoWorkshops/bioinfoTools/pattern.html

Dr. Horng's group develops systems to determine correlations of protein motifs [Horng et al 2003, 2004]. The knowledge of protein motif/domain sharing is important in finding biological functions of proteins and is useful in analyzing the evolution in the human genome or other genomes. They use PIR-NREF database and PROSITE database as data source. Then apply data mining to discover the occurrence correlations of motif in protein sequences.

Dr. Hu's group applies an iterative optimization method to find motifs based on object functions [Hu 1999]. With the motif finding algorithm, they use multiple objective functions and an improved stochastic iterative sampling strategy to perform combinatorial motif analysis; they use inductive learning algorithm to analyze potential motif combinations [Hu 2000].

Gene Finding/Annotation

For a general survey, readers can consult [Mathe et al 2002; Rogic et al 2001].

Dr. Ch'ang's group has developed a complexity reduction algorithm for sequence analysis (CRASA) that enables direct alignment of cDNA sequences to the genome [Chuang et al 2003b]. This method features a progressive data structure in hierarchical orders to facilitate a fast and efficient search mechanism. With its large-scale processing capability, CRASA can be used as a robust tool for genome annotation with high accuracy by matching the EST sequences precisely to the genomic sequences.

Protein Structure Prediction

For a general survey, readers can consult [Schonbrun 2002].

Dr. Hwang and Dr. Lyu's group have done a lot of works on protein structure prediction. They use many statistical model (such as support vector machine and evolutionary methods) and different measures (such as entropy or free energy) to predict protein folding procedure and side chain conformation [Chan et al 2003; Fan et al 1999; Yang et al 2002; Yu et al 2003a]. In particular, they find that protein structures can be determined by observing their disulfide-biding patterns [Chuang et al 2003a]. Since disulfide bond is only formed between a pair of Cysteine, so they use support vector machine to characterize the bonding state between amino acids [Chen et al 2004].

Dr. Hsu's group develops a knowledge-based approach for protein secondary struc-

structure prediction [Wu et al 2004]. The knowledge base contains small peptide fragments along with their structure information. They define a quantitative measure *M*, called the match rate, which represents the amount of structure information that a target protein can extract from the knowledge base. With the knowledge-based approach, they propose a hybrid prediction method as follows: predetermine a cutoff threshold value *M** for match rate; if the match rate of a target protein is greater than *M**, they use the extracted information to make the prediction; otherwise, they adopt a popular machine learning approach, such as PHD or PSIPRED. Because the approach seems to be complimentary to the machine learning ones, when compared to popular PHD or PSIPRED, it is likely that similar edge would prevail even after the machine learning approaches are later improved.

Protein Structure Identification

NMR is a powerful tool in protein structure identification. An important stage of protein structure determination by using NMR is protein backbone resonance assignment (or *backbone assignment* for short). NMR spectral data is usually transferred into spin systems. A spin system contains the chemical shifts of atoms within a residue. To perform backbone assignment, one usually needs to know the (partial) sequential order of spin systems; the procedure of figuring out such an order is called *connectivity determination*. Much work has been done for backbone assignment if a good connectivity determination result is given. However, due to the nature of NMR experiments, the spectral data usually contains noises, which makes connectivity determination difficult. Most researchers focus on backbone assignment and assume that a good connectivity determination result is available. Little has been done dealing with noisy data.

Dr. Hsu's group has proposed an iterative algorithm that can handle both connectivity determination and backbone assignment [Hsu et al 2004]. They solve the two problems in a relax fashion, and use a heuristic maximum independent set algorithm to perform backbone assignment.

Literature Mining

Information extraction over biomedical literatures is an emerging area in bioinformatics. A lots of papers has been accumulated in PubMed in which much information is presented in unstructured form, i.e. in natural language. Biomedical information extraction results can facilitate biologists' experimental design and database curation. The first step of information extraction is recognizing biological named entities or keywords, like gene name and protein name, on literatures. The next step is to identify relationships between these named entities. Some relations like protein-protein inter-

action, gene products-function and gene-disease are most popular subjects.

Although extracting information from unstructured text can facilitate biologists' experiment, a new challenge is how to convert information into knowledge and how to manage the knowledge. To cope with overwhelming data and texts biologists require a well design knowledge management system than traditional databases. Most knowledge management systems consist of two parts, ontology and knowledge application systems. Ontology represents knowledge in a structured fashion and knowledge application systems utilize ontology to help biologists query, update, and analyze data. Dr. Chiang's group develops ontology-based text mining system [Chiang & Yu 2003; Chiang et al 2004]. They extract gene-gene relations by aligning sentences. Dr. Chen's group extracts protein names by using both rule-based and corpus-based approaches, and then find relations between proteins by observing their collocations [Hou & Chen 2002]. Dr. Hsu's group develops a knowledge base, InfoMap, for proteomic named entity recognition (NER) and for named entity relationship recognition (NERR) from free text [Lin et al 2004]. The latter is critical for literature search. In NER, one needs to recognize protein names, genes and diseases. Morphological features and head noun features are stored in InfoMap for machine training and unknown word matching. For NERR, people are interested in events such as protein-protein interaction, protein-subcellular locations and so on. These systems help to semi-automatically acquire knowledge in InfoMap for building a biological question answering system.

Reference

Booth, K. and Lucker, G. Testing for the consecutive ones property, interval graphs, and graph planarity using PQ-tree algorithms. *J. of Computer and System Science* 13, 335-379, (1976).

Chan, C.C., Lyu, P.C. and Hwang, J.K. Computation of the Protein Structure Entropy and its Applications to Protein Folding Processes. *J. Chin. Chem. Soc.*, 50, 677-684, (2003)

Chao, K.M. Calign: Aligning Sequences with Restricted Affine Gap Penalties. *Bioinformatics*, 15: 298-304, (1999)

Chen, Y.C., Lin, Y.S., Lin, C.J., and Hwang, J.K. Prediction of the Bonding States of Cysteines using the Support Vector Machines Based on Multiple Feature Vectors and Cysteine State Sequences. *Proteins: Structure, Function and Genetics*, (2004) (in press).

Chen, Y.Y.L., Lu, S.H., Shih, S.C.E., and Hwang, M.J. Single Nucleotide Polymorphism Mapping Using Genome-Wide Unique Sequences. *Genome Research* 12:

1106-1111, (2002)

Chiang, J.H. and Yu, H.C. MeKE: Discovering the Functions of Gene Products from Biomedical Literature via Sentence Alignment. *Bioinformatics*, 19 (11) ,1417-1422, (2004)

Chiang, J.H., Yu, H.C. and Hsu, H.J. GIS: A Biomedical Text-Mining System for Gene Information Discovery. *Bioinformatics*, 20(1),120-121, (2003b)

Chuang, A., Chen, C. Y., Yang, J.M., Lyu, P. C. and Hwang, J.K. The detection of protein structural similarity using disulfide-binding patterns. *Proteins: Structure, Function, and Genetics*, vol. 53, pp. 1-5, (2003)

Chuang, T.J., Lin, W.C., Lee, H.C., Wang, C.W., Hsiao, K.L., Wang, Z.H., Shieh, D., Lin, S.C. and Ch'ang, L.Y. A Complexity Reduction Algorithm for Analysis and Annotation of Large Genomic Sequences. *Genome Research*, Vol. 13, No. 2, pp. 313-322, (2003)

Fan, Z.Z., Hwang, J.-K. and Warshel, A. Using simplified protein representation as a reference potential for all-atom calculations of folding free energy. *Theor Chem Acc*, 103 (1999) 1, 77-80, (1999)

Horng, J.T. and Cho, W.F. Predicting Regulatory Elements in Repetitive Sequences Using Transcription Factor Binding Sites. *Electronic Journal of Biotechnology*, Vol. 3, No. 3, (2000)

Horng, J.T. and Huang, H.D. Mining putative Regulatory Elements in promoter Regions of Saccharomyces cerevisiae. *In Silicon Biology*, 2, pp. 0-11, (2002a)

Horng, J.T., Huang, H.D., Lin, F.M. and Wu, L.C. The Repetitive Sequence Database and Mining Putative Regulatory Elements in Gene Promoter Regions. *Journal of Computational Biology*, Vol. 9, Issue 4, pp. 621-640, (2002b)

Horng, J.T., Huang, H.D., Wang, S.H., Lin, F.M. and Hwang, J.K. Computing Motif Correlations in Proteins. *Journal of Computational Chemistry*, Vol. 24, Issue 16, pp. 2032-2043, (2003)

Horng, J.T., Hwang, H.D. and Fang, S.F. Discovering Common Structural Motifs of Ribosomal RNA Secondary Structures. to Appear in *Journal of Bioinformatics and Bioengineering*, (2004a)

Hou, W.J. and Chen, H.H. Extracting Biological Keywords from Scientific Text. *Proceedings of 13th International Conference on Genome Informatics 2002*, December 16-18 2002, Tokyo, Japan, 2002, 571-573, (2002)

Hsu, F.R. and Chen, J.F. Aligning ESTs to Genome Using Multi-Layer Unique Makers. *Proc. Of 2003 IEEE Computer Society Bioinformatics Conference*, (2003)

Hsu, W.L., Chang, J.M., Chou, W.C., Chen, J.B., Wu, K.P., Sung, T.Y., Chang, C.F., Wu, W.J. and Huang, T.H. An Iterative Relaxation Technique for the NMR Backbone Assignment Problem. *BIBE*, (2004)

Hu,Y.J. Detecting Motifs from Sequences. *Proceedings of the 16th International Conference on Machine Learning (ICML-99)*, p181-190, (1999a)

Hu,Y.J. Combinatorial Motif Analysis and Hypothesis Generation on a Genomic Scale. *Bioinformatics*, Vol 16, p222-232, Oxford Univeristy Press, March, (2000a)

Huang, X. and Chao, K.M. A Generalized Global Alignment Algorithm. *Bioinformatics*, 19: 228-233, (2003)

Kao, M.Y. Tree construction and evolution trees. *SIAM Journal on Computing*, 27(6): p. 1592-1616, (1998)

Kim, J. and Warnow, T. Tutorial on Phylogenetic Tree Estimation. in *Intelligent Systems on Molecular Biology*. (1999)

Lassmann, T. and Sonnhammer, E.L.L. Quality assessment of multiple alignment programs. *FEBS Letters*, 529: p. 126?30, (2002)

Lin, Y.F., Chou, W.C., Wu, K.P., Sung, T.Y. and Hsu, W.L. InfoMap: A Framework for Integrated Proteomic Knowledge Base. *IPC, (*2004)

Lin, Y.L., Jiang, T. and Chao, K.M. Efficient Algorithm for Locating the Length-Constrained Heaviest Segments, with Applications to Biomolecular Sequence Analysis. *Journal of Computer and System Sciences (JCSS)*, 65: 570-586, (2002)

Lin, Y.L., Huang, X., Jiang, T. and Chao, K.M. MAVG: Locating Non-Overlapping Maximum Average Segments in a Given Sequence. *Bioinformatics*, 19: 151-152, (2003a)

Lin, Y.L. and Hsu, T.S. Efficient Algorithms for Descendent Subtrees Comparison of Phylogenetic Trees with Applications to Co-evolutionary Classifications in Bacterial Genome. *The 14th Annual International Symposium on Algorithms and Computation (ISAAC'03), Springer Verlag, Lecture Notes in Computer Science 2906*, pp 339-351. Kyoto, Japan, December 15-17, (2003b)

Lu, H.I., Goldwasser, M.H. and Kao, M.Y. Linear-Time Algorithms for Computing Maximum-Density Sequence Segments with Bioinformatics Applications. *Journal of Computer and System Sciences*, accepted, (2003a)

Lu, W.F. and Hsu, W.L. A Test for Interval Graphs on Noisy Data. *Lecture Notes in Computer Science* 2647, 195-208, (2003b)

Lu, W.F. and Hsu, W.L. A Test for the Consecutive Ones Property on Noisy Data -

Application to Physical Mapping and Sequence Assembly," *Journal of Computational Biology* 10(5), 709-735, (2004)

Lu, C.L., Tang, C.Y. and Lee, C.T. The Full Steiner Tree Problem in Phylogeny. *Theoretical Computer Science*, 306, 55-67, (2003c)

Lu, F.C., Tsai, Y.T. and Tang, C.Y. An Efficient External Sorting Algorithm. *Information Processing Letters*, Vol. 75, Issue: 4, pp. 159-163, (2000)

Mathe, C., Sagot, M.F., Schiex, T. and Rouze, P. Current methods of gene prediction, their strengths and weaknesses. *Nucleic Acids Research*, 2002. 30: p. 4103-4117.

Notredame, C. Recent progresses in multiple sequence alignment: a survey. *Pharmacogenomics*, 3(1): p. 131- 144, (2002)

Rogic, S., Mackworth, A.K. and Ouellette, B.F.F. Evaluation of gene finding programs. *Genome Research*, 11: p. 817-832, (2001)

Schonbrun, J., Wedemeyer, W.J. and Baker, D. Protein structure prediction in 2002. *Current Opinion in Structural Biology*, 12: p. 348-354, (2002)

Shih, C.C.A. and Li, W.H. GS-Aligner: A Novel Tool for Aligning Genomic Sequences Using Bit-Level Operations. *Mol. Biol. Evol.*, vol. 20, pp.1299-1309, (2003)

Shih, S.C.E. and Hwang, M.J. Protein structure comparison by probability-based matching of secondary structure elements. *Bioinformatics* 19: 735-741, (2003)

Tang, C.Y., Lu, C.L., Chang, M.D.T., Tsai Y.T, Sun, Y.J., Chao, K.M., Chang, J.M., Chiou, Y.H., Wu, C.M., Chang, H.T. and Chou, W.I. A Constrained Multiple Sequence Alignment Tool Development and its Application to RNase Family Alignment. *Journal of Bioinformatics and Computational Biology*, 1: 267-287, (2003)

Thompson, J.D., Plewniak, F. and Poch, O. A comprehensive comparison of multiple sequence alignment programs. *Nucleic Acids Research*, 1999. 27: p. 2682-2690.

Tsai, Y.T., Lu, C.L., Yu, C.T., and Huang, Y.P. MuSiC: A Tool for Multiple Sequence Alignment with Constraints. *Bioinformatics*, (2004)

Wu, B.Y., Chao, K.M., and Tang, C.Y. Approximation and Exact Algorithms for Constructing Minimum Ultrametric Trees from Distance Matrices. *Journal of Combinatorial Optimization*. Vol. 3, pp. 199-211, (1999)

Wu, K.P., Lin, H.N., Sung, T.Y. and Hsu, W.L. A New Similarity Measure among Protein Sequences. *Proceedings of 2nd IEEE Computer Society Bioinformatics (CSB 2003)*, pp. 347-352, August, (2003)

Wu, K.P., Lin, H.N., Chang, J.M., Sung, T.Y. and Hsu,W.L. PROSP-2D: A Hybrid Protein Secondary Structure Prediction Algorithm Based on a Knowledge-Based

Approach. *Taiwan-German Bioinformatics Conference*,(2004)

Yang, J.M., Tsai, C.H., Hwang, M.J., Tsai, H.K., Hwang, J.K. and Kao, C.Y. GEM: a Guassian Evolutionary Method for predicting protein Side-chain conformations. *Protein Science*, vol. 11, no. 8, pp. 1897-1907, (2002)

Yu, C.S., Wang, J.Y., Yang, J.M., Lyu, P.C., Lin, C.J., Hwang, J.K. Fine-grained Protein Fold Assignment by Support Vector Machines using generalized peptide Coding Schemes and jury voting from multiple parameter sets. *Proteins: Structure, Function, and Genetics*, vol. 50, pp. 531-536, (2003)