



POSTER ABSTRACTS – updated 11/14/06

Exploring Inconsistencies in Genome-Wide Protein Function Annotations

Presenter: Carson Andorf, Iowa State University

Author(s): Carson Andorf, Dr. Drena Dobbs, Dr. Vasant Honavar

Abstract:

Incorrectly annotated sequence data are becoming more commonplace as databases increasingly rely on automated techniques for annotation. Hence, there is an urgent need for computational methods for checking consistency of such annotations against independent sources of evidence and detecting potential annotation errors. We show how a machine learning approach designed to automatically predict a protein's Gene Ontology (GO) functional class can be employed to identify potential gene annotation errors. In a set of 211 previously annotated mouse protein kinases, we found that greater than 95% of the GO annotations returned by AmiGO appear to be inconsistent with the UniProt functions assigned to their human counterparts. In contrast, 97% of the predicted annotations generated using a machine learning approach were consistent with the UniProt annotations of the human counterparts, as well as with available annotations for these mouse protein kinases in the Mouse Kinome database. We conjecture that most of our predicted annotations are, therefore, correct and suggest that the machine learning approach developed here could be routinely used to detect potential errors in GO annotations generated by high-throughput gene annotation projects.

Phylogenetic Analysis of Teeth Decaying Oral Bacteria Through CLUSTAL W

Presenter: Abdul Arif Khan, Department of Microbiology, College of Life Sciences, Cancer Hospital & Research Institute

Author(s): Abdul Arif Khan

Abstract:

Various oral bacteria have suspected role in dental caries causation. Bacterial isolates from oral bacteria has been confirmed many times for their potential to produce acid, because acid production is cardinal for the development of dental caries. However the role of specific bacterial species or group of bacteria in dental caries is still controversial. Some researcher proposed the Streptococcus mutans as a major cause of dental caries, while some proposes role of other bacteria in this event. During present study, total sixty

one different bacterial strains were chosen for phylogenetic analysis. Total twelve Streptococcus sobrinus strains, thirteen strains of S. mutans, ten strains of S. salivarius, six strains of S. mitis, one strain of S. gordonii, three strains of S. sanguis, three strains of S. sanguinis, one strain of Uncultured Enterococci, one strain of S. italicus, seven strains of Actinomyces viscosus, two strains of Actinobacillus actinomycetemcomitans, and two strains of Lactobacillus acidophilus were analyzed. 16s RNA sequences of each strains was retrieved through Genebank and their multiple sequence alignment was performed through clustal W. Multiple sequence alignment revealed that Streptococcus salivarius accession AY581142 and Actinomyces viscosus accession no. M33908 are more related to each other, rather than their other strain of same genera. So, multiple sequence alignment can answer various enigma related to involvement of oral bacteria in dental caries causation, it can also give clue about the involvement of mixed bacteria in dental caries, as similarities between bacteria can also inform us about any methods to counteract these all caries bacteria by only a single drug by finding out their similarities and differences.

Using GoogleRanks™ for Phage Phylogeny

Presenter: Annalinda Arroyo, San Diego State University, Department of Mathematics and Statistics

Author(s): Annalinda Arroyo, Chrystian Irazoque

Abstract:

Using protein similarity, we define an approach to molecular based phylogeny for bacteriophages. The phylogeny uses phage-phage distances obtained by weighted averaging of the protein-protein distances between every pair of proteins. The weights in the averaging use GoogleRanks™ of the associated random walks. We expect to find an interesting and significant depiction of the relatedness between bacteriophage species.

Modeling Mixtures of Cells using Flow Cytometry

Presenter: Mike J Boedigheimer, Amgen

Author(s): MJ Boedigheimer

Abstract:

Flow Cytometry has become a standard technique for measuring physical attributes of single cells in a suspended cell mixture such as size, granularity and other indirectly labeled attributes by virtue of fluorescent labeled tags. By using different labels multiple attributes can be measured simultaneously allowing different subsets of cells to be studied independently. Currently, this is common done using a process of gating. Despite

being commonly used, gating has some well-known problems, such as being inefficient and prone to artifacts. Here we introduce a multivariate non-gating technique that accomplishes the same goals as gating while eliminating many of its weaknesses and an application to automate the analysis. The results are as good as expert manual gating with greatly reduced effort and provides for more robust statistical inference.

Information Retrieval, Question-Answering, Machine Learning, and Concept Recognition in TREC Genomics 2006

Presenter: J. Gregory Caporaso, University of Colorado Health Sciences Center, Dept. of Biochemistry and Molecular Genetics

Author(s): J. Gregory Caporaso, William A. Baumgartner, Jr., Hyunmin Kim, Zhiyong Lu, Helen L. Johnson, Olga Medvedeva, Anna Lindemann, Lynne Fox, Elizabeth K. White, K. Bretonnel Cohen, and Lawrence Hunter

Abstract:

TREC Genomics 2006 presented a challenge to automatically identify document passages that answer questions on twenty-seven topics from a collection of 162,259 full-text biomedical journal articles. Questions were derived from actual information needs of biomedical researchers, and performance was based on human evaluation of the retrieved passages. The Center for Computational Pharmacology approached this task by generating a candidate result set which was subsequently expanded (to improve recall) and then pruned (to improve precision). First, we created a Lemur search index and converted questions into term-expanded queries. Pseudo-relevance feedback was employed to expand search results. Next, from a pool of zone-filtered documents, we further expanded the result set using an LSA-like approach. In the final, false-positive eliminating step, we used naïve Bayesian classifiers trained on human-labeled data with features including bigrams, normalized semantic concepts, and OpenDMAP-style pattern matches. Three separate experiments allow us to compare the utility of the different techniques in this task.

Glycosylation Site Prediction Using Machine Learning Approaches

Presenter: Cornelia Caragea, Iowa State University

Author(s): C. Caragea, J. Sinapov, A. Silvescu, D. Dobbs and V. Honavar

Abstract:

Protein glycosylation is one of the most complex and important post-translational modifications (e.g. some proteins cannot fold unless they are glycosylated). It is a site-specific, enzymatic process of addition of saccharides to proteins, occurring on S and T

(O-Linked), N (N-Linked) and W residues (C-Linked). In this work, we address the problem of predicting glycosylation sites using different machine learning approaches applied to sequences for which different alternative representations are used. We investigate it on glycosylation data at O-GlycBase, a resource containing experimentally verified glycosylation sites. We build models on windows of different lengths centered at S, T, N, and W. Because our dataset is large and unbalanced, we use an ensemble classifier approach: we train a bag of classifiers, with each being trained on a balanced fraction of the data. Three types of classifiers were used: Support Vector Machine with different kernels: 0/1 String kernel, BLOSUM62 Substitution Matrix kernel, and Polynomial kernel, Naïve Bayes and Decision Tree. Our ensemble classifier predicted O-Linked sites with 0.89 accuracy, 0.57 correlation coefficient (CC), 0.65 precision, 0.62 recall, N-Linked sites with 0.93 accuracy, 0.76 CC, 0.75 precision, 0.86 recall, and C-Linked sites with 0.78 accuracy, 0.63 CC, 0.65 precision, 0.98 recall.

Structure-Activity Relationship Analysis of Chemical Compounds with Antioxidant Activity

Presenter: Weiguo Fan, Kent State University

Author(s): Weiguo Fan, Xin Lin, Boren Lin, Johnnie Baker, Chun-Che Tsai

Abstract:

A topological approach is used for structure-activity relationship (SAR) analysis of chemical compounds with antioxidant activity through structural activity maps (SAMs). SAMs are graphical maps plotting a molecular descriptor such as NAB (number of non-hydrogen atoms and bonds in a molecule) or MSI (molecular similarity index) against biological activity such as antioxidant activity. NAB or MSI is used to quantify the chemical structures. A program called TOPSIM was designed to find the maximal common substructure (MaCS) of a pair of molecules, each represented as a two-dimensional (2D) labeled graph. The molecular similarity index (MSI) and the topological distance (TD) define the similarity/dissimilarity of compounds. The algorithm uses a subgraph isomorphism approach to finding the maximum common subgraph (MCSG). TD and MSI calculated based on MaCS are used to build SAMs. SAMs provide a very efficient way of representing and visualizing SAR information in a chemical database. SAMs also allow comparison of grouped topological isomers and exploration of important trends in activity and sites of modification through structural orderings. A structural ordering is a set of structurally related compounds. It is utilized in this study to examine the effects of systematic modification of compounds on their biological activities.

A Customizable Evaluation Platform for Biomedical Information Extraction Systems

Presenter: Graciela Gonzalez, Arizona State University / School of Computing and Informatics, Dept of Biomedical Informatics

Author(s): Graciela Gonzalez, Anthony Gitter, Craig Teegarden, Chitta Baral

Abstract:

The ever growing number of biomedical articles published has created a dire need to access information hidden in text. Many biomedical information extraction systems have been developed to extract protein-protein interactions and other relationships such as gene-disease or gene-drug interactions. It is difficult to compare the performance of the extraction strategies because there is no standard for evaluation of such systems. In general, precision and recall are the performance measures used, but it is meaningless to compare these values when the tests performed vary so greatly. Developers make different decisions for critical aspects, and render the comparison of systems practically impossible, even when using the same articles and facts. We created a web-based platform that allows collaboration in extraction system evaluation and in creating "gold standards" that can be shared by multiple users. The platform allows evaluators to load PubMed abstracts or other references while viewing the extracted facts. The software is openly available and intuitive to use. Using the evaluation platform, developers are able to more accurately compare their system with competing systems, and identify weaknesses with greater ease. It is also a suitable platform for evaluating different systems, as is done for shared-tasks challenges like Biocreative and TREC.

New Method to De Novo Identify Repetitive Structures in Large Genomes

Presenter: Wanjun Gu, Department of Biochemistry and Molecular Genetics, University of Colorado Health Sciences Center

Author(s): Wanjun Gu, Dale J Hedges, Mark A Batzer, David D Pollock

Abstract:

The annotation of repeat structure in newly sequenced eukaryotic genomes can be difficult because of the lack knowledge of repeat consensus sequence in RepBase and the large amount of genome data that needs to be processed and compared. Here, we introduce a novel heuristic approach to de novo identify the repetitive sequence in large genomes, which avoids similarity searches and sequence alignments, the most two time-consuming steps in comparative analysis. We use the word counts in the genome to get the excessive observed words and then cluster them to get the oligonucleotide excess probability clouds, or "P clouds". After construction, P clouds were mapped back onto the genome, and regions of high P cloud density were identified as repetitive segments. This method is at least 10 times faster than current approaches for whole genome analysis. Comparing with the annotation of this method in human genome to RepeatMasker annotation, this method can mask most of the repeat elements

RepeatMasker identified. Analysis of simulated genome with short Alu-segments shows our method have higher sensitivity to annotate short segments of known repeat elements than RepeatMasker. This method is useful in comparative analysis of eukaryotic genomes and de novo repeat analysis in new genomes.

Comparative Studies of Elastin: Genomics of a Highly Repetitive Protein

Presenter: David He, University of Toronto

Author(s): David He, Fred W. Keeley, John Parkinson

Abstract:

Elastin is a polymeric, self-assembling, extracellular protein with remarkable durability, capable of withstanding billions of stretch-and-recoil cycles over decades without significant loss of performance or mechanical failure. As a result, elastin is an especially promising candidate for the development of novel biomaterials with wide ranging applications. However, while it is known that elastin's physical properties depend on its unusual sequence and domain structure, there is no clear understanding of how the sequence of elastin directly impacts these properties. The most significant challenge in unraveling these sequence-function relationships is the difficulty of analyzing elastin's repetitive yet highly variable sequence using traditional sequence analysis methods. We have applied an innovative application of tools previously developed for the study of protein interaction on a high quality, in-house elastin dataset, and have identified both common and species-specific 'core' structural elements within the elastin sequences. We further exploited this dataset to delineate a potential evolutionary path of elastin exons, suggesting duplicated regions which may have influenced the evolutionary-functional relationship of the elastin protein. Our methods have proven useful for the analysis of elastin sequences, and can be readily modified for other similarly repetitive sequences.

An Isometric Strength Testing Device for Use With Elderly: Validation Compared With Isokinetic Measures

Presenter: Alexander T. Hutchison, University of Houston

Author(s): Alexander T. Hutchison, Mark S.F. Clarke

Abstract:

Objective: To estimate the concurrent validity of a new isometric strength testing device, the Leg Press Sled (LPS), using a criterion measure of isokinetic strength among a frail elderly population.

Design: Prospective validation study.

Participants: Eleven elderly subjects (81.7 ± 7.0 yrs) from an assisted living facility.
Measurements: Measurements of isometric leg strength using the LPS were compared with measurements of isokinetic leg strength using the Biodex System 3, as a reference standard. For isometric measures, a sample of elderly subjects (mean 81.7 ± 7.0 yrs) was measured at knee angles 60° and 90° .

Results: Significant relationships ($R^2 > 0.50$, $P < .05$) between lower limb strength measures performed using the LPS and the Biodex System 3 were observed at both angles measured.

Conclusion: The LPS is a valid measure of leg strength in elderly populations. The device is portable, inexpensive, easy to use, and specifically suited for special-needs populations such as the home bound elderly due to its supine mode of operation.

Analysis of Ethnomedical plants with a potential for drug discovery - A biomedical informatics approach

Presenter: Beatrice Kilel, George Mason University

Author(s): Beatrice Kilel

Abstract:

Little information is currently known on some of the ethnomedical plants that have a potential for drug discovery. Hoodia cactus plant has recently captured headlines as an important resource with the potential as a biological pill for the obese based on the ingredient P57 that is known to suppress hunger. Some of the ethnomedical plants have been known to have a potential to cure some of the ailments that are not easy to cure using current drugs such as cancer. In order to ensure that these plants (such as *Prunus africana*, *Warburgia salutaris*) are not depleted, there has to be a way to regulate the harvesting regimes. Some of these resources have over time been pirated and patented in foreign countries without any profits to the rightful countries of origin.

Prunus africana is an ethnomedical plant that is traditionally used in Africa to treat ailments like chest pain, malaria and fevers. It has also been sought by large pharmaceutical companies for the manufacture of products to treat prostate gland hypertrophy (enlarged prostate gland) and benign prostatic hyperplasia (BPH). Bark extracts are pulverized and incorporated into capsules and sold under various trade names, including Pygenil, produced in Italy, and Tadenan, produced in France.

Warburgia salutaris is an important antimicrobial for treating yeast, fungal, bacterial and protozoal infections. It has been determined that the leaves and bark of this plant contains numerous drimane sesquiterpenoids, including warburganal and polygodial.

Interesting results were found in these less studied plants, and their potential for drug discovery. By having a more centralized repository to query from and make more scientific discoveries, more knowledge can be obtained about these ethnomedical resources.

Interpreting Gene Expression Data using Protein Interaction Networks

Presenter: Sonia Leach, University of Colorado, Pharmacology

Author(s): Sonia Leach

Abstract:

Protein interaction networks are a graphical means of representing the growing amount of information describing potential interaction between genes. Nodes in the graph correspond to genes and edges between the nodes represent a combination of physical, genetic, biochemical, functional, evolutionary or computational evidence of interaction between the corresponding genes. The challenge is to integrate the sources of interaction evidence while remaining sensitive to their relative reliability and coverage. We present methods which assign a reliability score to each evidence source and compute a probabilistic consensus likelihood of interaction. We demonstrate how the resulting consensus probabilities can be used to build visualization tools for the downstream analysis of gene expression data. Networks containing high confidence interactions provide a functional context for the genes and facilitate global interpretation of the observed changes in gene expression.

Protein 3D Structure Classification Using Distance Images and Needleman-Wunsch Algorithm

Presenter: Sherif H. El Meligy, Teacher Assistant

Author(s): Sherif H. El Meligy, Osama badawy Samah Sonbol

Abstract:

Since the 3D structure of a protein determines its function, the protein structural identification and classification is very important to biologists. Many protein representations and descriptors have been proposed. On the early days, the representation of protein structure to be used on classification consists of the position and the intra-molecule distance for all atoms consisting of protein. This representation had critical weak points due to high computational complexity and memory usage. In this paper we use an improved method by considering the position of alpha carbons (protein back bone) instead of using all atoms, Distance images are generated from the 3D coordinates of the alpha carbon atoms; then we apply a new technique to extract the information encoded in these images using block labeling and string alignment. Distance matrices or images are capable of representing specific protein structural topologies, and similar proteins in the same family tend to have similar distance matrices. A part of the SCOP database, which provides a structural classification of the proteins, is used as the ground truth in order to evaluate the classification accuracy of the proposed system; also a comparison is made

between our system and other methods for further evaluation. The experimental results show that the proposed system achieves more than 99 percent classification accuracy.

Spatio-Temporal Modelling of Biochemical Pathways: Introducing Cell++

Presenter: John Parkinson, Hospital for Sick Children / University of Toronto

Author(s): Chris Sanford, Matthew Yip and John Parkinson

Abstract:

High throughput genomic technologies are leading to the generation of vast amounts of data on cellular components, detailing their expression, interactions and organization within biochemical pathways. To understand how these relationships result in a functional pathway, a variety of computational tools have been proposed that attempt to simulate the behavior of the molecular components. In general, these tools tend to neglect the spatial organization of molecules, typically treating the system as a homogenous mixture of components. However, there is increasing evidence that spatial factors such as the co-localization of components have the potential to significantly influence pathway function and efficiency. Here we present Cell++, a novel spatio-temporal modeling platform that performs three-dimensional simulations of biochemical pathways.

Combining a cellular automata engine with Brownian dynamics, Cell++ is capable of simulating the bulk properties of large quantities of small molecules (e.g. pyruvate), while simultaneously allowing larger molecules such as enzymes to be treated as more complex entities. Applying Cell++ to the study of metabolic pathways, we demonstrate how the spatial organization of enzymes can alter pathway efficiency and control the production of substrate intermediates, features consistent with the phenomenon of metabolic channeling. Further details of Cell++ can be found at:

<http://www.compsysbio.org/CellSim/>.

The Conservation and Evolutionary Modularity of Metabolic Networks

Presenter: Jose M. Peregrin-Alvarez, Hospital for Sick Children University of Toronto

Author(s): Jose M. Peregrin-Alvarez, John Parkinson

Abstract:

Comparative genomics of metabolic pathways across genomes produces invaluable information on their evolution, potential drug targets, and other biotechnological applications. Here we investigate the level of conservation of metabolic enzymes and pathways, and the extent to what orthologs of proteins involved in metabolic pathways are co-occurring together across genomes (referred as evolutionary modularity). We also analyse from a topological and evolutionary point of view the network of interactions

where metabolic enzymes participate. We found that metabolic enzymes are highly conserved but not their participation in pathways. We generated a map of metabolic pathways that represents the similarity of metabolic pathway profiles and the extent of metabolic pathway similarity across taxa. Domain- and taxa-specific adaptations are identified and described. In addition, we found that metabolic pathways do not have a high degree of evolutionary modularity, although their modularity becomes more evident when metabolic pathways are analysed in the context of individual taxa. Metabolic network reconstruction also revealed a scale-free topology. Finally, for the very first time, we reveal a core metabolic network highly conserved across all kingdoms of life which potentially represents an ancestral state of the metabolic network.

Characterization of Binding Site of Cyclin-Dependent Kinase (CDK9) by Homology Modeling, Molecular Docking and Pharmacophore Identification

Presenter: Amresh Prakash, Jawaharlal Nehru University

Author(s): Amresh Prakash, Madhu Chopra

Abstract:

School of Information Technology, Jawaharlal Nehru University, New Delhi-110067, India. Abstract: Cyclin-dependent kinase 9 (CDK9) has significant role in muscle differentiation, neuron, astrocytes maturation and is one of the causal factor of heart hypertrophy¹. The present work describes the structural requirements for designing inhibitors for CDK9, which is an attempt towards Insilico drug discovery. As the 3D structure of CDK9 is not known till date, we built a model structure of CDK9 by homology modeling with template 1GZ8, 1V0O and 1H4L, which have share sequence identity more than 40% with the target. Pharmacophores were generated on the basis of known inhibitors of CDK2 and CDK5 using the tool Catalyst/HipHop. The pharmacophores, in general, were found to be characterized of four features, namely, two hydrogen bond acceptors, one hydrogen bond donor and one aromatic ring. We conducted virtual screening² using compound libraries - NCI and MayBridge to generate potential CDK9 inhibitors. The initial high numbers of hits generated by 3D database search were further searched for drug likeliness properties using CAS SciFinder, which finally gave four lead molecules as CDK9 inhibitors. Keywords: CDK 9, heart hypertrophy, homology modeling, Pharmacophore, Virtual screening, Catalyst/HipHop

A Computational Approach to the Evolution of Cytochrome P450 Structures

Presenter: Gowri Shankar, School of Information Technologies, University of Sydney

Author(s): Gowri Shankar, Michael Charleston, Michael Murray, David Hibbs

Abstract:

Cytochrome P450 (CYP) is an important enzyme in metabolizing drugs in humans and other species. This study aims at the evolution of this enzyme: How the structures have changed over time and between species and the significant changes in the drug binding site, which determines the function of the protein. We constructed evolutionary trees relating all 14 CYP proteins and used them to compare structures and sequences in a sound statistical framework. Structural differences were compared with the evolutionary distances between species. In the binding site of these proteins, there is a conserved pattern (F-G-X-G-X-H/R-X-C-L/I-G) with a histidine to arginine change in mammals from archaea and bacteria at position 6. There are substantial changes in the structure among species and there is less change in the binding site, though the function of the proteins changes substantially with time. Many hydrophobic residues are present in the helices of these proteins; these residues have a greater influence on the ligand and have a weak interaction with the ligand and the heme present in the CYPs. In summary, we note a remarkable consistency of some sites over 2 billion years, but clear evolutionary changes particular to some species.

The BioCyc Collection of Pathway/Genome Databases

Presenter: Alexander G. Shearer, SRI International

Author(s): Alexander G. Shearer, Ron Caspi, Carol A. Fulcher, Pallavi Kaipa, Peter D. Karp

Abstract:

The BioCyc collection of Pathway/Genome Databases (PGDBs) provides integrated representations of pathway and genome information for more than 200 organisms, of which most are microbes. Most BioCyc PGDBs were computationally derived from annotated genomes using the MetaCyc database (DB), which describes more than 800 experimentally determined metabolic pathways from more than 700 organisms. MetaCyc is a highly curated, literature-based DB of metabolic pathways and enzymes that provides a quality reference for pathway prediction. It is also an encyclopedic reference source for metabolic engineering and for other studies of metabolism. The computationally generated BioCyc PGDBs include predicted metabolic pathways as well as predicted fillers of holes in those metabolic pathways. BioCyc data are encoded using the Pathway Tools ontology, which facilitates the representation of complex biological knowledge with high fidelity. Pathway Tools provides a variety of visualization and analysis capabilities. Those to be presented here include its new comparative pathway analysis capabilities and its many data access mechanisms. Pathway Tools also has the ability to

automatically generate metabolic map diagrams for each organism in the BioCyc collection. These diagrams can be used for analysis of omics datasets and can be enlarged to produce publication-quality metabolic wall charts.

Ultrametric Structure Exhibited in Phage Proteins

Presenter: Chad Wagner, San Diego State University

Author(s): Chad Wagner

Abstract:

When phage proteins are compared to an ultrametric model, their phylogenetic structure adheres closely to the model for small neighborhoods. Statistical analysis of their deviations from ultrametricity reveals an interesting small scale structure and shows how this structure fits together into a larger geometric hierarchy.

Linguistic Features of Conflict and Support Statements in Biomedical Papers

Presenter: Elizabeth K. White, University of Colorado Health Sciences Center

Author(s): Elizabeth K. White

Abstract:

Scientific authors often note controversy in their field of research by way of introduction or discussion. Automatic recovery of conflicting and supporting evidence for a hypothesis can generate a bird's eye view of a field of study. Formal writing suggests several common rhetorical ways to express degree of agreement, and these can provide the basis for extracting support and conflict statements from scientific text. I analyzed six hundred support and conflict statements from biological papers to derive linguistic cues indicating corroboration or disagreement. I present a taxonomy of how these cues combine to form canonical conflict and support statements, and evaluate searches based on these cues against a test set of papers from the TREC 2006 corpus.

Alu-enhanced Sequence Diversity in the Human Genome

Presenter: Hong Xue, Department of Biochemistry, HKUST

Author(s): Siu-Kin NG, Hong Xue

Abstract:

Identifying features shaping the architecture of sequence variations is important for understanding genome evolution and mapping disease loci. In this study, high-resolution scanning of Alu-centered alignments of the human genome sequences has revealed a striking elevation of the frequency of single nucleotide polymorphisms inside Alu sequences (Alu-SNP) compared to their background sequence. This enhancement in SNP density is evident for all 24 chromosomes. Reduced levels of Alu-SNPs in the sex chromosomes, especially in the non-recombining region of the Y chromosome, are consistent with recombination events playing an important role in the enhancement. Variations in Alu-SNPs among the HapMap populations of European ancestry (CEU), Han Chinese from Beijing (CHB), Japanese from Tokyo (JPT), and Yoruba from Ibadan, Nigeria (YRI) indicate that Alu-SNPs provide useful sequence markers, in addition to the Alu-insertion polymorphisms themselves, for the delineation of human genome evolution. That Alu-SNP levels are highest in the youngest Alu-Y, intermediate in the Alu-S of intermediate age, and lowest in the oldest Alu-J is consistent with the occurrence of not only genetic drift but also natural selection on the Alu-SNPs. Such evolutionary selection in turn suggests that Alu-SNPs might include potential sites of disease association, and therefore deserve detailed investigation.