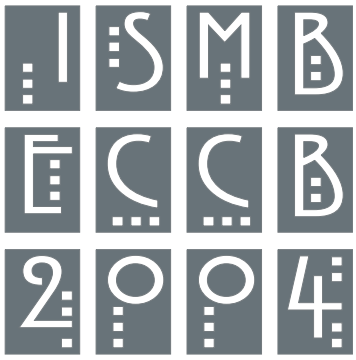
⁞⁞ PAPER ⁞⁞ PRESENTATIONS



Paper Presentations

Keynote Lecture

Sunday, 1 August **0900 - 0950**
Clyde Auditorium

Leroy Hood

Presentation 1

Sunday, 1 August **1010 - 1040**
Clyde Auditorium

IMGT/JunctionAnalysis: the first tool for the analysis of the immunoglobulin and T cell receptor complex V-J and V-D-J JUNCTIONs
Mehdi Yousfi Monod, Véronique Giudicelli, Denys Chaume and Marie-Paule Lefranc

Motivation: To create the enormous diversity of 1012 immunoglobulins (IG) and T cell receptors (TR) per individual, very complex mechanisms occur at the DNA level: the combinatorial diversity results from the junction of the variable (V), diversity (D) and joining (J) genes; the N-diversity represents the addition at random of nucleotides not encoded in the genome; and somatic hypermutations occur in IG rearranged sequences. The accurate annotation of the junction between V, D, J genes in rearranged IG and TR sequences represents therefore a huge challenge by its uniqueness and complexity. We developed IMGT/JunctionAnalysis to analyse automatically in detail the IG and TR junctions, according to the IMGT Scientific chart rules, based on the IMGT-ONTOLOGY concepts.

Results: IMGT/JunctionAnalysis is the first tool for the detailed analysis of the IG and TR complex V-J and V-D-J JUNCTION(s). It delimits, at the nucleotide level, the genes resulting from the combinatorial diversity. It identifies accurately the D genes in the junctions of IG heavy (IGH), TR beta (TRB) and delta (TRD) chains. It delimits the palindromic P-REGION(s) and the N-REGION(s) resulting from the N-diversity. It evaluates the number of somatic hypermutations for each gene, within the JUNCTION. IMGT/JunctionAnalysis is capable of analysing, in a single run, an unlimited number of junctions from the same species (currently human or mouse) and from the same locus.

Availability: IMGT/JunctionAnalysis is available from the IMGT Home page at http://imgt.cines.fr

Presentation 2

Sunday, 1 August **1010 - 1040**
Lomond Auditorium

The Gene Ontology Categorizer
Cliff A. Joslyn , Susan M. Mniszewski, Andy Fulmer and Gary Heaton

Summary: The Gene Ontology Categorizer, developed jointly by the Los Alamos National Laboratory and Procter & Gamble Corp., provides a capability for the categorization task in the Gene Ontology (GO): given a list of genes of interest, what are the best nodes of the GO to summarize or categorize that list? The motivating question is from a drug discovery process, where after some gene expression analysis experiment, we wish to understand the overall effect of some cell treatment or condition by identifying 'where' in the GO the differentially expressed genes fall: 'clustered' together in one place? in two places? uniformly spread throughout the GO? 'high', or 'low'? In order to address this need, we view bio-ontologies more as combinatorially structured databases than facilities for logical inference, and draw on the discrete mathematics of finite partially ordered sets (posets) to develop data representation and algorithms appropriate for the GO. In doing so, we have laid the foundations for a general set of methods to address not just the categorization task, but also other tasks (e.g. distances in ontologies and ontology merger and exchange) in both the GO and other bio-ontologies (such as the Enzyme Commission database or the MEdical Subject Headings) cast as hierarchically structured taxonomic knowledge systems.

Presentation 3

Sunday, 1 August **1110 - 1140**
Clyde Auditorium

Tracking repeats using significance and transitivity
Radek Szklarczyk and Jaap Heringa

Motivation: Internal repeats in coding sequences correspond to structural and functional units of proteins. Moreover, duplication of fragments of coding sequences is known to be a mechanism to facilitate evolution. Identification of repeats is crucial to shed light on the function and structure of proteins, and explain their evolutionary past. The task is difficult because during the course of evolution many repeats diverged beyond recognition.

Results: We introduce a new method TRUST, for ab initio determination of internal repeats in proteins. It provides an improvement in prediction quality as compared to alternative state-of-the-art methods. The increased sensitivity and accuracy of the method is achieved by exploiting the concept of transitivity of alignments. Starting from significant local sub-optimal alignments, the application of transitivity allows us to (1) identify distant repeat homologues for which no alignments were found; (2) gain confidence about consistently well-aligned regions; and (3) recognize and reduce the contribution of non-homologous repeats. This re-assessment step enables us to derive a virtually noise-free profile representing a generalized repeat with high fidelity. We also obtained superior specificity by employing rigid statistical testing for self-sequence and profile-sequence alignments. Assessment was done using a database of repeat annotations based on structural superpositioning. The results show that TRUST is a useful and reliable tool for mining tandem and non-tandem repeats in protein sequence databases, capable of predicting multiple repeat types with varying intervening segments within a single sequence.

Availability: The TRUST server (together with the source code) is available at http://ibivu.cs.vu.nl/programs/trustwww

Presentation 4

Sunday, 1 August **1110 - 1140**
Lomond Auditorium

HyBrow: a prototype system for computer-aided hypothesis evaluation
S. A. Racunas, N. H. Shah, I. Albert and N. V. Fedoroff

Motivation: Experimental design, hypothesis-testing and model-building in the current data-rich environment require the biologists' to collect, evaluate and integrate large amounts of information of many disparate kinds. Developing a unified framework for the representation and conceptual integration of biological data and processes is a major challenge in bioinformatics because of the variety of available data and the different levels of detail at which biological processes can be considered.

Results: We have developed the HyBrow (Hypothesis Browser) system as a prototype bioinformatics tool for designing hypotheses and evaluating them for consistency with existing knowledge. HyBrow consists of a modeling framework with the ability to accommodate diverse biological information sources, an event-based ontology for representing biological processes at different levels of detail, a database to query information in the ontology and programs to perform hypothesis design and evaluation. We demonstrate the HyBrow prototype using the galactose gene network in Saccharomyces cerevisiae as our test system, and evaluate alternative hypotheses for consistency with stored information.

Availability: www.hybrow.org

Presentation 5

Sunday, 1 August **1140 - 1210**
Clyde Auditorium

Efficient approximations for learning phylogenetic HMM models from data
Vladimir Jojic, Nebojsa Jojic, Chris Meek, Dan Geiger, Adam Siepel, David Haussler and D. Heckerman

Motivation: We consider models useful for learning an evolutionary or phylogenetic tree from data consisting of DNA sequences corresponding to the leaves of the tree. In particular, we consider a general probabilistic model described in Siepel and Haussler that we call the phylogenetic-HMM model which generalizes the classical probabilistic models of Neyman and Felsenstein. Unfortunately, computing the likelihood of phylogenetic-HMM models is intractable. We consider several approximations for computing the likelihood of such models including an approximation introduced in Siepel and Haussler, loopy belief propagation and several variational methods.

Results: We demonstrate that, unlike the other approximations, variational methods are accurate and are guaranteed to lower bound the likelihood. In addition, we identify a particular variational approximation to be best - one in which the posterior distribution is variationally approximated using the classic Neyman-Felsenstein model. The application of our best approximation to data from the cystic fibrosis transmembrane conductance regulator gene region across nine eutherian mammals reveals a CpG effect.

Presentation 6

Sunday, 1 August **1140 - 1210**
Lomond Auditorium

Mining MEDLINE for implicit links between dietary substances and diseases
Padmini Srinivasan and Bisharah Libbus

Motivation: Text mining systems aim at knowledge discovery from text collections. This work presents our text mining algorithm and demonstrates its use to uncover information that could form the basis of new hypotheses. In particular, we use it to discover novel uses for Curcuma longa, a dietary substance, which is highly regarded for its therapeutic properties in Asia.

Results: Several disease were identified that offer novel research contexts for curcumin. We analyze select suggestions, such as retinal diseases, Crohn's disease and disorders related to the spinal cord. Our analysis suggests that there is strong evidence in favor of a beneficial role for curcumin in these diseases. The evidence is based on curcumin's influence on several genes, such as COX-2, TNF-alpha, JNK, p38 MAPK and TGF-beta. This research suggests that our discovery algorithm may be used to suggest novel uses for dietary and pharmacological substances. More generally, our text mining algorithm may be used to uncover information that potentially sheds new light on a given topic of interest.

Availability: Contact authors.

Presentation 7

Sunday, 1 August **1210 - 1230**
Clyde Auditorium

MUSCLE: Low-complexity multiple sequence alignment with T-Coffee accuracy
Robert Edgar

We describe MUSCLE, a new program for creating multiple alignments of protein sequences. MUSCLE achieves the highest score so far reported on the BAliBASE benchmark, with average accuracy statistically indistinguishable from T-Coffee. MUSCLE aligns 5,000 sequences of average length 350 in 7 minutes on a current desktop computer, requiring less time than all other tested methods, including MAFFT. We also introduce PREFAB, a new multiple alignment benchmark. PREFAB results confirm that MUSCLE and T-Coffee produce, on average, the most accurate alignments, with 6% more positions correctly aligned than ClustalW. Software, source code and test data is freely available at: http://www.drive5.com/muscle.

Presentation 8

Sunday, 1 August **1210 - 1240**
Lomond Auditorium

Protein names precisely peeled off free text

Motivation: Automatically identifying protein names from the scientific literature is a pre-requisite or the increasing demand in data-mining this wealth of information. Existing approaches are based on dictionaries, rules and machine-learning. Here, we introduced a novel system that combines a pre-processing dictionary- and rule-based filtering step with several separately trained support vector machines (SVMs) to identify protein names in the MEDLINE abstracts.

Results: Our new tagging-system NLProt is capable of extracting protein names with a precision (accuracy) of 75% at a recall (coverage) of 76% after training on a corpus, which was used before by other groups and contains 200 annotated abstracts. For our estimate of sustained performance, we considered partially identified names as false positives. One important issue frequently ignored in the literature is the redundancy in evaluation sets. We suggested some guidelines for removing overly inadequate overlaps between training and testing sets. Applying these new guidelines, our program appeared to significantly out-perform other methods tagging protein names. NLProt was so successful due to the SVM-building blocks that succeeded in utilizing the local context of protein names in the scientific literature. We challenge that our system may constitute the most general and precise method for tagging protein names.

Availability:
http://cubic.bioc.columbia.edu/services/nlprot/

Presentation 9

Sunday, 1 August **1230 - 1250**
Clyde Auditorium

PROBCONS: probabilistic consistency-based multiple alignment of amino acid sequences
Chuong Do, Michael Brudno, Serafim Batzoglou

Obtaining an accurate multiple alignment of protein sequences is often difficult when amino acid percent identity is low. In this paper, we present PROBCONS, a practical tool for protein multiple sequence alignment, based on an algorithm that combines HMM-derived posterior probabilities with consistency-based alignment techniques. On the BAliBASE benchmark alignment database, PROBCONS demonstrates a statistically significant improvement in accuracy compared to several leading alignment programs while maintaining practical running times. Source code of the program is freely available under the GNU Public License at http://probcons.stanford.edu/.

Presentation 10

Sunday, 1 August **1240 - 1300**
Lomond Auditorium

False annotations of proteins: automatic setection via keyword-based clustering
Noam Kaplan, Michal Linial

Computational protein annotation methods occasionally introduce errors. False-positive (FP) errors are annotations that are mistakenly associated with a protein. Such false annotations introduce errors that may spread into databases through similarity with other proteins. We present a protein-clustering method that enables automatic separation of FP from true-positive hits. The method is based on the combination of each protein's annotations. Using a test set of all PROSITE signatures that are marked as FPs, we show that the method successfully separates FPs in 70% of the cases. Automatic detection of FPs may greatly facilitate the manual validation process and increase annotation sensitivity.

Presentation 11

Sunday, 1 August **1450 - 1520**
Clyde Auditorium

Reconstructing tumor amplisomes
Benjamin J. Raphael and Pavel A. Pevzner

Motivation: Duplication of genomic sequences is a common phenomenon in tumor cells. While many duplications associated with tumors have been identified (e.g. via techniques such as CGH), both the organization of the duplicated sequences and the process that leads to these duplications are less clear. One mechanism that has been observed to lead to duplication is the extraction of DNA from the chromosomes and aggregation of this DNA into small, independently replicating linear or circular DNA sequences (amplisomes). Parts of these amplisomes may later be reinserted back into the main chromosomes leading to duplication. Although amplisomes are known to play an important role in tumorigenesis, their architecture and even size remain largely unknown.

Results: We reconstruct the structure of tumor amplisomes by analyzing duplications in the tumor genome. Our approach relies on recently generated data from End Sequence Profiling (ESP) experiments, which allow us to examine the fine structure of duplications in a tumor on a genome-wide scale. Using ESP data, we formulate the Amplisome Reconstruction Problem, describe an algorithm for its solution, and derive a putative architecture of a tumor amplisome that is the source for duplicated material in the MCF7 breast tumor cell line.

Presentation 12

Sunday, 1 August **1450 - 1520**
Lomond Auditorium

Splice site identification by idlBNs
Robert Castelo and Roderic Guigó

Motivation: Computational identification of functional sites in nucleotide sequences is at the core of many algorithms for the analysis of genomic data. This identification is based on the statistical parameters estimated from a training set. Often, because of the huge number of parameters, it is difficult to obtain consistent estimators. To simplify the estimation problem, one imposes independent assumptions between the nucleotides along the site. However, this can potentially limit the minimum value of the estimation error.

Results: In this paper, we introduce a novel method in the context of identifying functional sites, that finds a reasonable set of independence assumptions supported by the data, among the nucleotides, and uses it to perform the identification of the sites by their likelihood ratio. More importantly, in many practical situations it is capable of improving its performance as the training sample size increases. We apply the method to the identification of splice sites, and further evaluate its effect within the context of exon and gene prediction.

Supplementary information: The datasets built specifically for this paper as well as the full set of results are available at
http://genome.imim.es/datasets/splidlbns2004

Presentation 13

Sunday, 1 August **1520 - 1550**
Clyde Auditorium

The cell graphs of cancer
Cigdem Gunduz , BülentYener and S. Humayun Gultekin

Summary: We report a novel, proof-of-concept, computational method that models a type of brain cancer (glioma) only by using the topological properties of its cells in the tissue image. From low-magnification (80¥) tissue images of 384 ¥ 384 pixels, we construct the graphs of the cells based on the locations of the cells within the images. We generate such cell graphs of 1000-3000 cells (nodes) with 2000-10 000 links, each of which is calculated as a decaying exponential function of the Euclidean distance between every pair of cells in accordance with the Waxman model. At the cellular level, we compute the graph metrics of the cell graphs, including the degree, clustering coefficient, eccentricity and closeness for each cell. Working with a total of 285 tissue samples surgically removed from 12 different patients, we demonstrate that the self-organizing clusters of cancerous cells exhibit distinctive graph metrics that distinguish them from the healthy cells and the unhealthy inflamed cells at the cellular level with an accuracy of at least 85%. At the tissue level, we accomplish correct tissue classifications of cancerous, healthy and non-neoplastic inflamed tissue samples with an accuracy of 100% by requiring correct classification for the majority of the cells within the tissue sample.

Presentation 14

Sunday, 1 August **1520 - 1550**
Lomond Auditorium

Improved techniques for the identification of pseudogenes
L. Coin and R. Durbin

Motivation: Pseudogenes are the remnants of genomic sequences of genes which are no longer functional. They are frequent in most eukaryotic genomes, and an important resource for comparative genomics. However, pseudogenes are often mis-annotated as functional genes in sequence databases. Current methods for identifying pseudogenes include methods which rely on the presence of stop codons and frameshifts, as well as methods based on the ratio of non-silent to silent nucleotide substitution rates (dN/dS). A recent survey concluded that 50% of human pseudogenes have no detectable truncation in their pseudo-coding regions, indicating that the former methods lack sensitivity. The latter methods have been used to find sets of genes enriched for pseudogenes, but are not specific enough to accurately separate pseudogenes from expressed genes.

Results: We introduce a program called pseudogene inference from loss of constraint (PSILC) which incorporates novel methods for separating pseudogenes from functional genes. The methods calculate the log-odds score that evolution along the final branch of the gene tree to the query gene has been according to the following constraints:
A neutral nucleotide model compared to a Pfam domain encoding model (PSILCnuc/dom);
A protein coding model compared to a Pfam domain encoding model (PSILCprot/dom).
Using the manual annotation of human chromosome 6, we show that both these methods result in a more accurate classification of pseudogenes than dN/dS when a Pfam domain alignment is available.

Availability: PSILC is available from http://www.sanger.ac.uk/Software/PSILC

Presentation 15

Sunday, 1 August **1550 - 1620**
Clyde Auditorium

Prediction of class I T-cell epitopes: evidence of presence of immunological hot spots inside antigens
K. N. Srinivasan, G. L. Zhang, A. M. Khan, J. T. August and V. Brusic

Motivation: Processing and presentation of major histocompatibility complex class I antigens to cytotoxic T-lymphocytes is crucial for immune surveillance against intracellular bacteria, parasites, viruses and tumors. Identification of antigenic regions on pathogen proteins will play a pivotal role in designer vaccine immunotherapy. We have developed a system that not only identifies high binding T-cell antigenic epitopes, but also class I T-cell antigenic clusters termed immunological hot spots.

Methods: MULTIPRED, a computational system for promiscuous prediction of HLA class I binders, uses artificial neural networks (ANN) and hidden Markov models (HMM) as predictive engines. The models were rigorously trained, tested and validated using experimentally identified HLA class I T-cell epitopes from human melanoma related proteins and human papillomavirus proteins E6 and E7.We have developed a scoring scheme for identification of immunological hot spots for HLA class I molecules, which is the sum of the highest four predictions within a window of 30 amino acids.

Results: Our predictions against experimental data from four melanoma-related proteins showed that MULTIPRED ANN and HMM models could predict T-cell epitopes with high accuracy. The analysis of proteins E6 and E7 showed that ANN models appear to be more accurate for prediction of HLA-A3 hot spots and HMM models for HLA-A2 predictions. For illustration of its utility we applied MULTIPRED for prediction of promiscuousT-cell epitopes in all four SARS coronavirus structural proteins.MULTIPRED predicted HLA-A2 and HLA-A3 hot spots in each of these proteins.

Presentation 16

Sunday, 1 August **1550 - 1620**
Lomond Auditorium

Exploiting conserved structure for faster annotation of non-coding RNAs without loss of accuracy
Zasha Weinberg and Walter L. Ruzzo

Motivation: Non-coding RNAs (ncRNAs) - functional RNA molecules not coding for proteins - are grouped into hundreds of families of homologs. To find new members of an ncRNA gene family in a large genome database, covariance models (CMs) are a useful statistical tool, as they use both sequence and RNA secondary structure information. Unfortunately, CM searches are slow. Previously, we introduced 'rigorous filters', which provably sacrifice none of CMs' accuracy, although often scanning much faster. A rigorous filter, using a profile hidden Markov model (HMM), is built based on the CM, and filters the genome database, eliminating sequences that provably could not be annotated as homologs. The CM is run only on the remainder. Some biologically important ncRNA families could not be scanned efficiently with this technique, largely due to the significance of conserved secondary structure relative to primary sequence in identifying these families. Current heuristic filters are also expected to perform poorly on such families.

Results: By augmenting profile HMMs with limited secondary structure information, we obtain rigorous filters that accelerate CM searches for virtually all known ncRNA families from the Rfam Database and tRNA models in tRNAscan-SE. These filters scan an 8 gigabase database in weeks instead of years, and uncover homologs missed by heuristic techniques to speed CM searches.

Availability: Software in development; contact the authors.

Supplementary information:
http://bio.cs.washington.edu/supplements/zasha-ISMB-2004
(Additional technical details on the method; predicted homologs.)

Presentation 17

Sunday, 1 August **1650 - 1720**
Clyde Auditorium

Exploring Williams-Beuren syndrome using myGrid
R. D. Stevens , H. J. Tipney, C. J. Wroe, T. M. Oinn, M. Senger, P. W. Lord, C. A. Goble, A. Brass and M. Tassabehji

Motivation: In silico experiments necessitate the virtual organization of people, data, tools and machines. The scientific process also necessitates an awareness of the experience base, both of personal data as well as the wider context of work. The management of all these data and the co-ordination of resources to manage such virtual organizations and the data surrounding them needs significant computational infrastructure support.

Results: In this paper, we show that myGrid, middleware for the Semantic Grid, enables biologists to perform and manage in silico experiments, then explore and exploit the results of their experiments. We demonstrate myGrid in the context of a series of bioinformatics experiments focused on a 1.5 Mb region on chromosome 7 which is deleted in Williams-Beuren syndrome (WBS). Due to the highly repetitive nature of sequence flanking/in the WBS critical region (WBSCR), sequencing of the region is incomplete leaving documented gaps in the released sequence. myGrid was used in a series of experiments to find newly sequenced human genomic DNA clones that extended into these 'gap' regions in order to produce a complete and accurate map of the WBSCR. Once placed in this region, these DNA sequences were analysed with a battery of prediction tools in order to locate putative genes and regulatory elements possibly implicated in the disorder. Finally, any genes discovered were submitted to a range of standard bioinformatics tools for their characterization. We report how myGrid has been used to create workflows for these in silico experiments, run those workflows regularly and notify the biologist when new DNA and genes are discovered. The myGrid services collect and co-ordinate data inputs and outputs for the experiment, as well as much provenance information about the performance of experiments on WBS.

Availability: The myGrid software is available via http://www.mygrid.org.uk

Presentation 18

Sunday, 1 August **1650 - 1720**
Lomond Auditorium

Functional inference from non-random distributions of conserved predicted transcription factor binding sites
Christoph Dieterich, Sven Rahmann and Martin Vingron

Motivation: Our understanding of how genes are regulated in a concerted fashion is still limited. Especially, complex phenomena like cell cycle regulation in multicellular organisms are poorly understood. Therefore, we investigated conserved predicted transcription factor binding sites (TFBSs) in man-mouse upstream regions of genes that can be associated to a particular cell cycle phase in HeLa cells. TFBSs were predicted from selected binding site motifs (represented by position weight matrices, PWMs) based on a statistical approach. A regulatory role for a transcription factor is more probable if its predicted TFBSs are enriched in upstream regions of genes, that are associated with a subset of cell cycle phases. We tested for this association by computing exact P-values for the observed phase distributions under the null distribution defined by the relative amount of conserved upstream sequence of genes per cell cycle phase. We considered non-exonic and 5¢-untranslated region (5¢-UTR) binding sites separately and corrected for multiple testing by taking the false discovery rate into account.

Results: We identified 22 non-exonic and 11 5¢-UTR significant PWM phase distributions although expecting one false discovery. Many of the corresponding transcription factors (e.g. members of the thyroid hormone/retinoid receptor subfamily) have already been associated with cell cycle regulation, proliferation and development. It appears that our method is a suitable tool for detecting putative cell cycle regulators in the realm of known human transcription factors.

Availability: Further details and supplementary data can be obtained from http://corg.molgen.mpg.de/cellcycle

Keynote Lecture

Sunday, 1 August **1720 - 1800**
Clyde Auditorium

Denis Noble

Keynote Lecture

Monday, 2 August **0830 - 0920**
Clyde Auditorium

Eric Green

Presentation 19

Monday, 2 August **0920 - 0940**
Clyde Auditorium

CHAINER: software for comparing genomes
Mohamed Ibrahim Abouelhoda, Enno Ohlebusch

Recently, software-tools for pairwise or multiple comparison of genomic sequences have gained an enormous importance in comparative genomics. Our novel software CHAINER can be used for several comparative tasks: (1) finding regions of high similarity (candidate regions of conserved synteny), (2) multiple global alignment of whole genomes, (3) comparison of multiple draft (or finished) genomes among themselves, and (4) cDNA/EST mapping. The software is available upon request.

Presentation 20

Monday, 2 August **0940 - 1010**
Clyde Auditorium

Genomic features in the breakpoint regions between syntenic blocks
Phil Trinh, Aoife McLysaght and David Sankoff

Motivation: We study the largely unaligned regions between the syntenic blocks conserved in humans and mice, based on data extracted from the UCSC genome browser. These regions contain evolutionary breakpoints caused by inversion, translocation and other processes.

Results: We suggest explanations for the limited amount of genomic alignment in the neighbourhoods of breakpoints. We discount inferences of extensive breakpoint reuse as artifacts introduced during the reconstruction of syntenic blocks. We find that the number, size and distribution of small aligned fragments in the breakpoint regions depend on the origin of the neighbouring blocks and the other blocks on the same chromosome. We account for this and for the generalized loss of alignment in the regions partially by artefacts due to alignment protocols and partially by mutational processes operative only after the rearrangement event. These results are consistent with breakpoints occurring randomly over virtually the entire genome.

Presentation 21

Monday, 2 August **0940 - 1010**
Lomond Auditorium

Predicting gene regulation by sigma factors in Bacillus subtilis from genome-wide data
M. J. L. de Hoon , Y. Makita, S. Imoto, K. Kobayashi, N. Ogasawara, K. Nakai and S. Miyano

Motivation: Sigma factors regulate the expression of genes in Bacillus subtilis at the transcriptional level. We assess the accuracy of a fold-change analysis, Bayesian networks, dynamic models and supervised learning based on coregulation in predicting gene regulation by sigma factors from gene expression data. To improve the prediction accuracy, we combine sequence information with expression data by adding their log-likelihood scores and by using a logistic regression model. We use the resulting score function to discover currently unknown gene regulations by sigma factors.

Results: The coregulation-based supervised learning method gave the most accurate prediction of sigma factors from expression data. We found that the logistic regression model effectively combines expression data with sequence information. In a genome-wide search, highly significant logistic regression scores were found for several genes whose transcriptional regulation is currently unknown. We provide the corresponding RNA polymerase binding sites to enable a straightforward experimental verification of these predictions.

Presentation 22

Monday, 2 August **1010 - 1040**
Clyde Auditorium

Using MoBIoS' scalable genome join to find conserved primer pair candidates between two genomes
Weijia Xu, Willard J. Briggs, Joanna Padolina, Ruth E. Timme, Wenguo Liu, C. Randal Linder and Daniel P. Miranker

Motivation: For the purpose of identifying evolutionary reticulation events in flowering plants, we determine a large number of paired, conserved DNA oligomers that may be used as primers to amplify orthologous DNA regions using the polymerase chain reaction (PCR).

Results: We develop an initial candidate set by comparing the Arabidopsis and rice genomes using MoBIoS (Molecular Biological Information System). MoBIoS is a metric-space database management system targeting life science data. Through the use of metric-space indexing techniques, two genomes can be compared in O(mlog n), where m and n are the lengths of the genomes, versus O(mn) for BLAST-based analysis. The filtering of low-complexity regions may also be accomplished by directly assessing the uniqueness of the region. We describe mSQL, a SQL extension being developed for MoBIoS that encapsulates the algorithmic details in a common database programming language, shielding end-users from esoteric programming.

Availability: Available upon request from authors.

Presentation 23

Monday, 2 August **1010 - 1040**
Lomond Auditorium

Deconvolving cell cycle expression data with complementary information
Ziv Bar-Joseph , Shlomit Farkash, David K. Gifford, Itamar Simon and Roni Rosenfeld

Motivation: In the study of many systems, cells are first synchronized so that a large population of cells exhibit similar behavior. While synchronization can usually be achieved for a short duration, after a while cells begin to lose their synchronization. Synchronization loss is a continuous process and so the observed value in a population of cells for a gene at time t is actually a convolution of its values in an interval around t. Deconvolving the observed values from a mixed population will allow us to obtain better models for these systems and to accurately detect the genes that participate in these systems.

Results: We present an algorithm which combines budding index and gene expression data to deconvolve expression profiles. Using the budding index data we first fit a synchronization loss model for the cell cycle system. Our deconvolution algorithm uses this loss model and can also use information from co-expressed genes, making it more robust against noise and missing values. Using expression and budding data for yeast we show that our algorithm is able to reconstruct a more accurate representation when compared with the observed values. In addition, using the deconvolved profiles we are able to correctly identify 15% more cycling genes when compared to a set identified using the observed values.

Availability: Matlab implementation can be downloaded from the supporting website http://www.cs.cmu.edu/~zivbj/decon/decon.html

Presentation 24

Monday, 2 August **1110 - 1140**
Clyde Auditorium

High density linkage disequilibrium mapping using models of haplotype block variation
G. Greenspan and D. Geiger

Motivation: The presence of millions of single nucleotide polymorphisms (SNPs) in the human genome has spurred interest in genetic mapping methods based on linkage disequilibrium. The recently discovered haplotype block structure of human variation promises to improve the effectiveness of these methods. A key difficulty for mapping techniques is the cost involved in separately identifying the haplotypes on each of an individual's chromosomes.

Results: We present a new approach for performing linkage disequilibrium mapping using high density haplotype or genotype data. Our method is based on a statistical model of haplotype block variation, which takes account of recombination hotspots, bottlenecks, genetic drift and mutation. We test our technique on two empirically determined high density datasets, attempting to recover the location of an SNP which was hidden and converted into phenotype information. We compare the results against a mapping method based on individual SNPs as well as a competing haplotype-based approach. We show that our strategy significantly outperforms these other approaches when used as a guide for resequencing and that it can also deal with both unphased genotype data and low penetrance diseases.

Availability: HaploBlock executables for Linux, Mac OS X and Sun OS, as well as user documentation, are available online at http://bioinfo.cs.technion.ac.il/haploblock/

Presentation 25

Monday, 2 August **1110 - 1140**
Lomond Auditorium

Statistical modeling of sequencing errors in SAGE libraries
Tim Beißbarth , Lavinia Hyde, Gordon K. Smyth, Chris Job, Wee-Ming Boon, Seong-Seng Tan, Hamish S. Scott and Terence P. Speed

Motivation: Sequencing errors may bias the gene expression measurements made by Serial Analysis of Gene Expression (SAGE). They may introduce non-existent tags at low abundance and decrease the real abundance of other tags. These effects are increased in the longer tags generated in Long-SAGE libraries. Current sequencing technology generates quite accurate estimates of sequencing error rates. Here we make use of the sequence neighborhood of SAGE tags and error estimates from the base-calling software to correct for such errors.

Results: We introduce a statistical model for the propagation of sequencing errors in SAGE and suggest an Expectation-Maximization (EM) algorithm to correct for them given observed sequences in a library and base-calling error estimates. We tested our method using simulated and experimental SAGE libraries. When comparing SAGE libraries, we found that sequencing errors can introduce considerable bias. High abundance tags may be falsely called as significantly differentially expressed, especially when comparing libraries with different levels of sequencing errors and/or of different size. Truly, differentially expressed tags have decreased significance as 'true'-tag counts are generally underestimated. This may alter if tags near the threshold of differential expression are called significant. Moreover, the number of different transcripts present in a library is overestimated as false tags are introduced at low abundance. Our correction method adjusts the tag counts to be closer to the true counts and is able to partly correct for biases introduced by sequencing errors.

Availability: An implementation using R is distributed as an R package. An online version is available at http://tagcalling.mbgproject.org

Presentation 26

Monday, 2 August **1140 - 1210**
Clyde Auditorium

Into the heart of darkness: large-scale clustering of human non-coding DNA
Gill Bejerano, David Haussler and Mathieu Blanchette

Motivation: It is currently believed that the human genome contains about twice as much non-coding functional regions as it does protein-coding genes, yet our understanding of these regions is very limited.

Results: We examine the intersection between syntenically conserved sequences in the human, mouse and rat genomes, and sequence similarities within the human genome itself, in search of families of non-protein-coding elements. For this purpose we develop a graph theoretic clustering algorithm, akin to the highly successful methods used in elucidating protein sequence family relationships.

The algorithm is applied to a highly filtered set of about 700 000 human-rodent evolutionarily conserved regions, not resembling any known coding sequence, which encompasses 3.7% of the human genome. From these, we obtain roughly 12 000 non-singleton clusters, dense in significant sequence similarities. Further analysis of genomic location, evidence of transcription and RNA secondary structure reveals many clusters to be significantly homogeneous in one or more characteristics. This subset of the highly conserved non-protein-coding elements in the human genome thus contains rich family-like structures, which merit in-depth analysis.

Availability: Supplementary material to this work is available at http://www.soe.ucsc.edu/~jill/dark.html

Presentation 27

Monday, 2 August **1140 - 1210**
Lomond Auditorium

Optimal robust non-unique probe selection using Integer Linear Programming
Gunnar W. Klau, Sven Rahmann, Alexander Schliep, Martin Vingron and Knut Reinert

Motivation: Besides their prevalent use for analyzing gene expression, microarrays are an efficient tool for biological, medical and industrial applications due to their ability to assess the presence or absence of biological agents, the targets, in a sample. Given a collection of genetic sequences of targets one faces the challenge of finding short oligonucleotides, the probes, which allow detection of targets in a sample. Each hybridization experiment determines whether the probe binds to its corresponding sequence in the target. Depending on the problem, the experiments are conducted using either unique or non-unique probes and usually assume that only one target is present in the sample. The problem at hand is to compute a design, i.e. a minimal set of probes that allows to infer the targets in the sample from the result of the hybridization experiment. If we allow to test for more than one target in the sample, the design of the probe set becomes difficult in the case of non-unique probes.

Results: Building upon previous work on group testing for microarrays, we describe the first approach to select a minimal probe set for the case of non-unique probes in the presence of a small number of multiple targets in the sample. The approach is based on an ILP formulation and a branch-and-cut algorithm. Our preliminary implementation greatly reduces the number of probes needed while preserving the decoding capabilities.

Availability: http://www.inf.fu-berlin.de/inst/ag-bio

Presentation 28

Monday, 2 August **1210 - 1230**
Clyde Auditorium

CIS: compound importance sampling method for protein-dna binding site p-value estimation
Yoseph Barash, Gal Elidan, Tommy Kaplan, Nir Friedman

Motivation: Transcription regulation involves binding of transcription factors to sequence-specific sites and controlling the expression of nearby genes. Given binding site models, one can scan the regulatory regions for putative binding sites and construct a genome-wide regulatory network. Several recent works demonstrated the importance of modeling dependencies between positions in the binding site. The challenge is to evaluate the statistical significance of binding sites using these models. Results: We present a general, accurate and efficient method for this task, applicable to any probabilistic binding site and background models. We demonstrate the accuracy of the method on synthetic and real-life data.

Availability: The algorithm used to compute the statistical significance of putative binding sites scores is available online at http://compbio.cs.huji.ac.il/CIS/ Contact: nir@cs.huji.ac.il

Presentation 29

Monday, 2 August **1210 - 1230**
Lomond Auditorium

TIDE - Terra Incognita Discovery Endeavor: comprehensive EST assignment to GeneCards genes
Maxim Shklar, Orit Shmueli, Liora Strichman-Almashanu, Michael Shmoish, Tsippy Iny-Stein, Marilyn Safran, Doron Lancet

The construction of a complete EST-based gene index has been an intricate task. We present TIDE, an automated system for associating each of the >5 million human ESTs with a known or de-novo defined gene. The pipeline is heavily based on existing GeneCards links to other EST-related resources. In a specific example, we were able to provide gene identities to an additional ~15,000 unassigned EST-based Affymetrix microarray probesets, a 50% increase relative to previous annotations. TIDE is expected to help complete a comprehensive EST-associated compendium of GeneCards genes.

Presentation 30

Monday, 2 August **1420 - 1450**
Clyde Auditorium

Reconstructing phylogeny by Quadratically Approximated Maximum Likelihood
M. D. Woodhams and M. D. Hendy

Summary: Maximum likelihood (ML) for phylogenetic inference from sequence data remains a method of choice, but has computational limitations. In particular, it cannot be applied for a global search through all potential trees when the number of taxa is large, and hence a heuristic restriction in the search space is required. In this paper, we derive a quadratic approximation, QAML, to the likelihood function whose maximum is easily determined for a given tree. The derivation depends on Hadamard conjugation, and hence is limited to the simple symmetric models of Kimura and of Jukes and Cantor. Preliminary testing has demonstrated the accuracy of QAML is close to that of ML.

Presentation 31

Monday, 2 August **1420 - 1450**
Lomond Auditorium

Finding disease specific alterations in the co-expression of genes
Dennis Kostka and Rainer Spang

Motivation: Standard analysis routines for microarray data aim at differentially expressed genes. In this paper, we address the complementary problem of detecting sets of differentially co-expressed genes in two phenotypically distinct sets of expression profiles.

Results: We introduce a score for differential co-expression and suggest a computationally efficient algorithm for finding high scoring sets of genes. The use of our novel method is demonstrated in the context of simulations and on real expression data from a clinical study.

Presentation 32

Monday, 2 August **1450 - 1520**
Clyde Auditorium

Phylogenomics and the number of characters required for obtaining an accurate phylogeny of eukaryote model species
Hernán Dopazo, Javier Santoyo and Joaquín Dopazo

Motivation: Through the most extensive phylogenomic analysis carried out to date, complete genomes of 11 eukaryotic species have been examined in order to find the homologous of more than 25 000 amino acid sequences. These sequences correspond to the exons of more than 3000 genes and were used as presence/absence characters to test one of the most controversial hypotheses concerning animal evolution, namely the Ecdysozoa hypothesis. Distance, maximum parsimony and Bayesian methods of phylogenetic reconstruction were used to test the hypothesis.

Results: The reliability of the ecdysozoa, grouping arthropods and nematodes in a single clade was unequivocally rejected in all the consensus trees. The Coelomata clade, grouping arthropods and chordates, was supported by the highest statistical confidence in all the reconstructions. The study of the dependence of the genomes' tree accuracy on the number of exons used, demonstrated that an unexpectedly larger number of characters are necessary to obtain robust phylogenies. Previous studies supporting ecdysozoa, could not guarantee an accurate phylogeny because the number of characters used was clearly below the minimum required.

Presentation 33

Monday, 2 August **1450 - 1520**
Lomond Auditorium

Partial Cox regression analysis for high-dimensional microarray gene expression data
Hongzhe Li and Jiang Gui

Motivation: An important application of microarray technology is to predict various clinical phenotypes based on the gene expression profile. Success has been demonstrated in molecular classification of cancer in which different types of cancer serve as categorical outcome variable. However, there has been less research in linking gene expression profile to censored survival outcome such as patients' overall survival time or time to cancer relapse. In this paper, we develop a partial Cox regression method for constructing mutually uncorrelated components based on microarray gene expression data for predicting the survival of future patients.

Results: The proposed partial Cox regression method involves constructing predictive components by repeated least square fitting of residuals and Cox regression fitting. The key difference from the standard principal components of Cox regression analysis is that in constructing the predictive components, our method utilizes the observed survival/censoring information. We also propose to apply the time-dependent receiver operating characteristic curve analysis to evaluate the results. We applied our methods to a publicly available dataset of diffuse large B-cell lymphoma. The outcomes indicated that combining the partial Cox regression method with principal components analysis results in parsimonious model with fewer components and better predictive performance. We conclude that the proposed partial Cox regression method can be very useful in building a parsimonious predictive model that can accurately predict the survival of future patients based on the gene expression profile and survival times of previous patients.

Availability: R codes are available upon request.

Presentation 34

Monday, 2 August **1520 - 1550**
Clyde Auditorium

Evolution of multi-domain proteins by gene fusion and fission
Sarah Kummerfeld, Christine Vogel, Martin Madera, Mary Pacold, Sarah Teichmann

During evolution genes may undergo recombination to produce more complex proteins by gene fusion. The reverse process generates multiple less complex proteins by fission. Considering proteins from 131 genomes, we established rates of gene fusion and fission. We found 2869 groups of multi-domain proteins that exist as single proteins in some organisms, and two or more smaller proteins in others. Applying maximum parsimony to species trees, we found fusion events were four times more common than fission. Analysing the functions of these proteins reveals that one third of split proteins are subunits of a single function complex, while two thirds have separate functions in the same pathway or as components of a multi-functional complex.

Presentation 35

Monday, 2 August **1520 - 1550**
Lomond Auditorium

Predicting genetic regulatory response using classification
Manuel Middendorf, Anshul Kundaje, Chris Wiggins, Yoav Freund and Christina Leslie

Motivation: Studying gene regulatory mechanisms in simple model organisms through analysis of high-throughput genomic data has emerged as a central problem in computational biology. Most approaches in the literature have focused either on finding a few strong regulatory patterns or on learning descriptive models from training data. However, these approaches are not yet adequate for making accurate predictions about which genes will be up- or down-regulated in new or held-out experiments. By introducing a predictive methodology for this problem, we can use powerful tools from machine learning and assess the statistical significance of our predictions.

Results: We present a novel classification-based method for learning to predict gene regulatory response. Our approach is motivated by the hypothesis that in simple organisms such as Saccharomyces cerevisiae, we can learn a decision rule for predicting whether a gene is up- or down-regulated in a particular experiment based on (1) the presence of binding site subsequences ('motifs') in the gene's regulatory region and (2) the expression levels of regulators such as transcription factors in the experiment ('parents'). Thus, our learning task integrates two qualitatively different data sources: genome-wide cDNA microarray data across multiple perturbation and mutant experiments along with motif profile data from regulatory sequences.We convert the regression task of predicting real-valued gene expression measurements to a classification task of predicting ++1 and -.1 labels, corresponding to up- and down-regulation beyond the levels of biological and measurement noise in microarray measurements. The learning algorithm employed is boosting with a margin-based generalization of decision trees, alternating decision trees. This large-margin classifier is sufficiently flexible to allow complex logical functions, yet sufficiently simple to give insight into the combinatorial mechanisms of gene regulation. We observe encouraging prediction accuracy on experiments based on the Gasch S. cerevisiae dataset, and we show that we can accurately predict up- and down-regulation on held-out experiments. We also show how to extract significant regulators, motifs and motif-regulator pairs from the learned models for various stress responses. Our method thus provides predictive hypotheses, suggests biological experiments, and provides interpretable insight into the structure of genetic regulatory networks.

Availability: The MLJava package is available upon request to the authors.

Supplementary: Additional results are available from http://www.cs.columbia.edu/compbio/geneclass

Presentation 36

Monday, 2 August **1620 - 1650**
Clyde Auditorium

A nucleotide substitution model with nearest-neighbour interactions
Gerton Lunter and Jotun Hein

Motivation: It is well known that neighbouring nucleotides in DNA sequences do not mutate independently of each other. In this paper, we introduce a context-dependent substitution model and derive an algorithm to calculate the likelihood of sequences evolving under this model. We use this algorithm to estimate neighbour-dependent substitution rates, as well as rates for dinucleotide substitutions, using a Bayesian sampling procedure. The model is irreversible, giving an arrow to time, and allowing the position of the root between a pair of sequences to be inferred without using out-groups.

Results: We applied the model upon aligned human-mouse non-coding data. Clear neighbour dependencies were observed, including 17-18-fold increased CpG to TpG/CpA rates compared with other substitutions. Root inference positioned the root halfway the mouse and human tips, suggesting an approximately clock-like behaviour of the irreversible part of the subsitution process.

Presentation 37

Monday, 2 August **1620 - 1650**
Lomond Auditorium

Robust inference of groups in gene expression time-courses using mixtures of HMMs
Alexander Schliep , Christine Steinhoff and Alexander Schönhuth

Motivation: Genetic regulation of cellular processes is frequently investigated using large-scale gene expression experiments to observe changes in expression over time. This temporal data poses a challenge to classical distance-based clustering methods due to its horizontal dependencies along the time-axis. We propose to use hidden Markov models (HMMs) to explicitly model these time-dependencies. The HMMs are used in a mixture approach that we show to be superior over clustering. Furthermore, mixtures are a more realistic model of the biological reality, as an unambiguous partitioning of genes into clusters of unique functional assignment is impossible. Use of the mixture increases robustness with respect to noise and allows an inference of groups at varying level of assignment ambiguity. A simple approach, partially supervised learning, allows to benefit from prior biological knowledge during the training. Our method allows simultaneous analysis of cyclic and non-cyclic genes and copes well with noise and missing values.

Results: We demonstrate biological relevance by detection of phase-specific groupings in HeLa time-course data. A benchmark using simulated data, derived using assumptions independent of those in our method, shows very favorable results compared to the baseline supplied by k-means and two prior approaches implementing model-based clustering. The results stress the benefits of incorporating prior knowledge, whenever available.

Availability: A software package implementing our method is freely available under the GNU general public license (GPL) at http://ghmm.org/gql

Supplementary information: Supplemental material can be found at http://algorithmics.molgen.mpg.de/ExpMix

Presentation 38

Monday, 2 August **1650 - 1720**
Clyde Auditorium

Transitions at CpG dinucleotides: are replication errors to blame?
Kateryna Makova, James Taylor

The sequenced human and chimpanzee genomes allow one to investigate mutations with high rates, since the divergence between the two genomes is low. We compared the rates of transitions at CpG dinucleotides among sex chromosomes and autosomes at noncoding sequences orthologous between human and chimpanzee. We discovered that in primates the transition rates at CpG dinucleotides are lowest for chromosome X, intermediate for autosomes, and highest for Y. Thus, these mutations primarily originate in the male germline which undergoes a larger number of DNA replications compared with the female germline. This suggests that transitions at CpG dinucleotides are replication-dependent.

Presentation 39

Monday, 2 August **1650 - 1720**
Lomond Auditorium

An information theory approach for validating clusters in microarray data
Sudhakar Jonnalagadda, Rajagopalan Srinivasan

Cluster validation is commonly used for evaluating the quality of partition produced by any clustering algorithm. In this paper, we present a novel method to assess the quality of clustering in gene expression data. In contrast to methods which are totally based on intra- and inter-cluster distances, our approach considers the dynamics and rearrangement of elements when a new cluster is introduced. Cluster quality is measured based on information change and the partition with the highest total information is selected. We illustrate the efficacy of the proposed method using two microarray datasets and two artificial datasets and discuss the advantages and limitations. **Key words:** Cluster validation, Gene Expression, Information Theory.

Supplementary information available at
http://cheed.nus.edu.sg/~chergs/ismb2004.htm

Keynote Lecture

Monday, 2 August **1720 - 1800**
Clyde Auditorium

Svante Pääbo

Keynote Lecture

Tuesday, 3 August **0830 - 0920**
Clyde Auditorium

Matthias Mann

Presentation 40

Tuesday, 3 August **0940 - 1010**
Clyde Auditorium

Analysis of domain correlations in yeast protein complexes
Doron Betel, Ruth Isserlin and Christopher W. V. Hogue

Motivation: A growing body of research has concentrated on the identification and definition of conserved sequence motifs. It is widely recognized that these conserved sequence and structural units often mediate protein functions and interactions. The continuing advancements in high-throughput experiments necessitate the development of computational methods to critically assess the results. In this work, we analyzed high-throughput protein complexes using the domain composition of their protein constituents. Domains that mediate similar or related functions may consistently co-occur in protein complexes.

Results: We analyzed Saccharomyces cerevisiae protein complexes from curated and high-throughput experimental datasets to identify statistically significant functional associations between domains. The resulting correlations are represented as domain networks that form the basis of comparison between the datasets, as well as to binary protein interactions. The results show that the curated datasets produce domain networks that map to known biological assemblies, such as ribosome, RNA polymerase, proteasome regulators, transcription initiation and histones. Furthermore, many of these domain correlations were also found in binary protein interactions. In contrast, the high-throughput datasets contain one large network of domain associations. High connectivity of RNA processing and binding domains in the high-throughput datasets reflects the abundance of RNA binding proteins in yeast, in agreement with a previous report that identified a nucleolar protein cluster, possibly mediated by rRNA, from these complexes.

Availability: The software is available upon request from the authors and is dependent on the NCBI C++ ++ toolkit.

Presentation 41

Tuesday, 3 August **0940 - 1010**
Lomond Auditorium

A neural-network-based method for predicting protein stability changes upon single point mutations
Emidio Capriotti, Piero Fariselli and Rita Casadio

Motivation: One important requirement for protein design is to be able to predict changes of protein stability upon mutation. Different methods addressing this task have been described and their performance tested considering global linear correlation between predicted and experimental data. Neither is direct statistical evaluation of their prediction performance available, nor is a direct comparison among different approaches possible. Recently, a significant database of thermodynamic data on protein stability changes upon single point mutation has been generated (ProTherm). This allows the application of machine learning techniques to predicting free energy stability changes upon mutation starting from the protein sequence.

Results: In this paper, we present a neural-network-based method to predict if a given mutation increases or decreases the protein thermodynamic stability with respect to the native structure. Using a dataset consisting of 1615 mutations, our predictor correctly classifies >80% of the mutations in the database. On the same task and using the same data, our predictor performs better than other methods available on the Web. Moreover, when our system is coupled with energy-based methods, the joint prediction accuracy increases up to 90%, suggesting that it can be used to increase also the performance of pre-existing methods, and generally to improve protein design strategies.

Availability: The server is under construction and will be available at http://www.biocomp.unibo.it

Presentation 42

Tuesday, 3 August **1010 - 1040**
Clyde Auditorium

Predicting protein-peptide interactions via a network-based motif sampler
David J. Reiss and Benno Schwikowski

Motivation: Many protein-protein interactions are mediated by peptide recognition modules (PRMs), compact domains that bind to short peptides, and play a critical role in a wide array of biological processes. Recent experimental protein interaction data provide us with an opportunity to examine whether we may explain, or even predict their interactions by computational sequence analysis. Such a question was recently posed by the use of random peptide screens to characterize the ligands of one such PRM, the SH3 domain.

Results: We describe a general computational procedure for identifying the ligand peptides of PRMs by combining protein sequence information and observed physical interactions into a simple probabilistic model and from it derive an interaction mediated de novo motif-finding framework. Using a recent all-versus-all yeast two-hybrid SH3 domain interaction network, we demonstrate that our technique can be used to derive independent predictions of interactions mediated by SH3 domains. We show that only when sequence information is combined with such all versus all protein interaction datasets, are we capable of identifying motifs with sufficient sensitivity and specificity for predicting interactions. The algorithm is general so that it may be applied to other PRM domains (e.g. SH2, WW, PDZ).

Availability: The Netmotsa software and source code, as part of a general Gibbs motif sampling library, are available at http://sf.net/projects/netmotsa

Presentation 43

Tuesday, 3 August **1010 - 1040**
Lomond Auditorium

Mining frequent patterns in protein structures: a study of protease families
Shann-Ching Chen and Ivet Bahar

Motivation: Analysis of protein sequence and structure databases usually reveal frequent patterns (FP) associated with biological function. Data mining techniques generally consider the physicochemical and structural properties of amino acids and their microenvironment in the folded structures. Dynamics is not usually considered, although proteins are not static, and their function relates to conformational mobility in many cases.

Results: This work describes a novel unsupervised learning approach to discover FPs in the protein families, based on biochemical, geometric and dynamic features. Without any prior knowledge of functional motifs, the method discovers the FPs for each type of amino acid and identifies the conserved residues in three protease subfamilies; chymotrypsin and subtilisin subfamilies of serine proteases and papain subfamily of cysteine proteases. The catalytic triad residues are distinguished by their strong spatial coupling (high interconnectivity) to other conserved residues. Although the spatial arrangements of the catalytic residues in the two subfamilies of serine proteases are similar, their FPs are found to be quite different. The present approach appears to be a promising tool for detecting functional patterns in rapidly growing structure databases and providing insights in to the relationship among protein structure, dynamics and function.

Availability: Available upon request from the authors.

Presentation 44

Tuesday, 3 August **1110 - 1140**
Clyde Auditorium

Automatic Quality Assessment of Peptide Tandem Mass Spectra
Marshall Bern, David Goldberg, W. Hayes McDonald and John R. Yates, III

Motivation: A powerful proteomics methodology couples high-performance liquid chromatography (HPLC) with tandem mass spectrometry and database-search software, such as SEQUEST. Such a set-up, however, produces a large number of spectra, many of which are of too poor quality to be useful. Hence a filter that eliminates poor spectra before the database search can significantly improve throughput and robustness. Moreover, spectra judged to be of high quality, but that cannot be identified by database search, are prime candidates for still more computationally intensive methods, such as de novo sequencing or wider database searches including post-translational modifications.

Results: We report on two different approaches to assessing spectral quality prior to identification: binary classification, which predicts whether or not SEQUEST will be able to make an identification, and statistical regression, which predicts a more universal quality metric involving the number of b- and y-ion peaks. The best of our binary classifiers can eliminate over 75% of the unidentifiable spectra while losing only 10% of the identifiable spectra. Statistical regression can pick out spectra of modified peptides that can be identified by a de novo program but not by SEQUEST. In a section of independent interest, we discuss intensity normalization of mass spectra.

Presentation 45

Tuesday, 3 August **1110 - 1140**
Lomond Auditorium

Striped sheets and protein contact prediction
Robert M. MacCallum

Motivation: Current approaches to contact map prediction in proteins have focused on amino acid conservation and patterns of mutation at sequentially distant positions. This sequence information is poorly understood and very little progress has been made in this area during recent years.

Results: In this study, an observation of 'striped' sequence patterns across b-sheets prompted the development of a new type of contact map predictor. Computer program code was evolved with an evolutionary algorithm (genetic programming) to select residues and residue pairs likely to make contacts based solely on local sequence patterns extracted with the help of self-organizing maps. The mean prediction accuracy is 27% on a validation set of 156 domains up to 400 residues in length, where contacts are separated by at least 8 residues and length/10 pairs are predicted. The retrospective accuracy on a set of 15 CASP5 targets is 27% and 14% for length/10 and length/2 predicted pairs, respectively (both using a minimum residue separation of 24). This compares favourably to the equivalent 21% and 13% obtained for the best automated contact prediction methods at CASP5. The results suggest that protein architectures impose regularities in local sequence environments. Other sources of information, such as correlated/compensatory mutations, may further improve accuracy.

Availability: A web-based prediction service is available at
http://www.sbc.su.se/~maccallr/contactmaps

Presentation 46

Tuesday, 3 August **1140 - 1210**
Clyde Auditorium

MSMS Peak Identification and its Applications
Navdeep Jaitly, Rachel Page-Belanger, Denis Faubert, Pierre Thibault, Paul Kearney

A peak detection algorithm for Tandem Mass Spectra is presented that scores a fragment using intensity and isotopic distribution. It classifies each fragment in a spectrum as noise or signal based on a maximum likelihood estimate derived from the distribution observed in a training set of 12,000 validated spectra. This is the largest such database known to the authors. We present three tools which apply this algorithm: the Quality Filter removes noisy spectra, Mod-Pro profiles modifications and amino acids in a sample and Spectrimilarity scores similarity of two spectra.

Presentation 47

Tuesday, 3 August **1140 - 1210**
Lomond Auditorium

SCOPEC: a database of protein catalytic domains
Richard A. George , Ruth V. Spriggs, JanetM. Thornton, Bissan Al-Lazikani and Mark B. Swindells

Motivation: Domains are the units of protein structure, function and evolution. It is therefore essential to utilize knowledge of domains when studying the evolution of function, or when assigning function to genome sequence data. For this purpose, we have developed a database of catalytic domains, SCOPEC, by combining structural domain information from SCOP, full-length sequence information from Swiss-Prot, and verified functional information from the Enzyme Classification (EC) database. Two major problems need to be overcome to create a database of domain-function relationships; (1) for sequences, EC numbers are typically assigned to whole sequences rather than the functional unit, and (2) The Protein Data Bank (PDB) structures elucidated from a larger multi-domain protein will often have EC annotation although the relevant catalytic domain may lie elsewhere.

Results: SCOPEC entries have high quality enzyme assignments; having passed both computational and manual checks. SCOPEC currently contains entries for 75% of all EC annotations in the PDB. Overall, EC number is fairly well conserved within a superfamily, even when the proteins are distantly related. Initial analysis is encouraging; suggesting that there is a 50:50 chance of conserved function in distant homologues first detected by a third iteration PSI-BLAST search. Therefore, we envisage that a knowledge-based approach to function assignment using the domain-EC relationships in SCOPEC will gain a marked improvement over this base line.

Availability: The SCOPEC database is a valuable resource in the analysis and prediction of protein structure and function. It can be obtained or queried at our website http://www.enzome.com

Presentation 48

Tuesday, 3 August **1210 - 1230**
Clyde Auditorium

Cross-species protein identification in proteomics via protein profiles
Patrick Lester, Christian Cole, Simon Hubbard

Protein identification from mass spectra of tryptic peptides relies on bioinformatic software to determine the most likely matching protein from a database that contains the protein sequence. However, enabling cross-species proteomics is equally important for the many species currently without sequenced genomes. We present a methodology to address this using profiles of related protein sequences against which to search. Using simulated data, we show that tryptic peptide conservation is enriched above random in these protein profiles, and a search algorithm developed from these shows an improvement over simple single-orthologue search methods.

Presentation 49

Tuesday, 3 August **1420 - 1450**
Clyde Auditorium

Assigning transmembrane segments to helices in intermediate-resolution structures
Angela Enosh, Sarel J. Fleishman, Nir Ben-Tal and Dan Halperin

Motivation: Transmembrane (TM) proteins that form a-helix bundles constitute approximately 50% of contemporary drug targets. Yet, it is difficult to determine their high-resolution (<4 Å) structures. Some TM proteins yield more easily to structure determination using cryo electron microscopy (cryo-EM), though this technique most often results in lower resolution structures, precluding an unambiguous assignment of TM amino acid sequences to the helices seen in the structure. We present computational tools for assigning the TM segments in the protein's sequence to the helices seen in cryo-EM structures.

Results: The method examines all feasible TM helix assignments and ranks each one based on a score function that was derived from loops in the structures of soluble a-helix bundles. A set of the most likely assignments is then suggested. We tested the method on eight TM chains of known structures, such as bacteriorhodopsin and the lactose permease. Our results indicate that many assignments can be rejected at the outset, since they involve the connection of pairs of remotely placed TM helices. The correct assignment received a high score, and was ranked highly among the remaining assignments. For example, in the lactose permease, which contains 12 TM helices, most of which are connected by short loops, only 12 out of 479 million assignments were found to be feasible, and the native one was ranked first.

Availability: The program and the non-redundant set of protein structures used here are available at http://www.cs.tau.ac.il/~angela

Presentation 50

Tuesday, 3 August **1420 - 1450**
Lomond Auditorium

Learning kernels from biological networks by maximizing entropy
Koji Tsuda and William Stafford Noble

Motivation: The diffusion kernel is a general method for computing pairwise distances among all nodes in a graph, based on the sum of weighted paths between each pair of nodes. This technique has been used successfully, in conjunction with kernel-based learning methods, to draw inferences from several types of biological networks.

Results: We show that computing the diffusion kernel is equivalent to maximizing the von Neumann entropy, subject to a global constraint on the sum of the Euclidean distances between nodes. This global constraint allows for high variance in the pairwise distances. Accordingly, we propose an alternative, locally constrained diffusion kernel, and we demonstrate that the resulting kernel allows for more accurate support vector machine prediction of protein functional classifications from metabolic and protein-protein interaction networks.

Availability: Supplementary results and data are available at noble.gs.washington.edu/proj/maxent

Presentation 51

Tuesday, 3 August **1450 - 1520**
Clyde Auditorium

Predicting protein folding pathways
Mohammed J. Zaki, Vinay Nadimpally, Deb Bardhan and Chris Bystroff

Summary: A structured folding pathway, which is a time ordered sequence of folding events, plays an important role in the protein folding process and hence, in the conformational search. Pathway prediction, thus gives more insight into the folding process and is a valuable guiding tool to search the conformation space. In this paper, we propose a novel 'unfolding' approach to predict the folding pathway. We apply graph-based methods on a weighted secondary structure graph of a protein to predict the sequence of unfolding events. When viewed in reverse this yields the folding pathway. We demonstrate the success of our approach on several proteins whose pathway is partially known.

Presentation 52

Tuesday, 3 August **1450 - 1520**
Lomond Auditorium

Protein network inference from multiple genomic data: a supervised approach
Y. Yamanishi, J.-P. Vert and M. Kanehisa

Motivation: An increasing number of observations support the hypothesis that most biological functions involve the interactions between many proteins, and that the complexity of living systems arises as a result of such interactions. In this context, the problem of inferring a global protein network for a given organism, using all available genomic data about the organism, is quickly becoming one of the main challenges in current computational biology.

Results: This paper presents a new method to infer protein networks from multiple types of genomic data. Based on a variant of kernel canonical correlation analysis, its originality is in the formalization of the protein network inference problem as a supervised learning problem, and in the integration of heterogeneous genomic data within this framework. We present promising results on the prediction of the protein network for the yeast Saccharomyces cerevisiae from four types of widely available data: gene expressions, protein interactions measured by yeast two-hybrid systems, protein localizations in the cell and protein phylogenetic profiles. The method is shown to outperform other unsupervised protein network inference methods. We finally conduct a comprehensive prediction of the protein network for all proteins of the yeast, which enables us to propose protein candidates for missing enzymes in a biosynthesis pathway.

Availability: Softwares are available upon request.

Presentation 53

Tuesday, 3 August **1520 - 1550**
Clyde Auditorium

A two-stage classifier for identification of protein-protein interface residues
Changhui Yan , Drena Dobbs and Vasant Honavar

Motivation: The ability to identify protein-protein interaction sites and to detect specific amino acid residues that contribute to the specificity and affinity of protein interactions has important implications for problems ranging from rational drug design to analysis of metabolic and signal transduction networks.

Results: We have developed a two-stage method consisting of a support vector machine (SVM) and a Bayesian classifier for predicting surface residues of a protein that participate in protein-protein interactions. This approach exploits the fact that interface residues tend to form clusters in the primary amino acid sequence. Our results show that the proposed two-stage classifier outperforms previously published sequence-based methods for predicting interface residues. We also present results obtained using the two-stage classifier on an independent test set of seven CAPRI (Critical Assessment of PRedicted Interactions) targets. The success of the predictions is validated by examining the predictions in the context of the three-dimensional structures of protein complexes.

Supplementary information:
http://www.public.iastate.edu/~chhyan/ISMB2004/list.html

Presentation 54

Tuesday, 3 August **1520 - 1550**
Lomond Auditorium

Inferring quantitative models of regulatory networks from expression data
I. Nachman, A. Regev and N. Friedman

Motivation: Genetic networks regulate key processes in living cells. Various methods have been suggested to reconstruct network architecture from gene expression data. However, most approaches are based on qualitative models that provide only rough approximations of the underlying events, and lack the quantitative aspects that are critical for understanding the proper function of biomolecular systems.

Results: We present fine-grained dynamical models of gene transcription and develop methods for reconstructing them from gene expression data within the framework of a generative probabilistic model. Unlike previous works, we employ quantitative transcription rates, and simultaneously estimate both the kinetic parameters that govern these rates, and the activity levels of unobserved regulators that control them. We apply our approach to expression datasets from yeast and show that we can learn the unknown regulator activity profiles, as well as the binding affinity parameters. We also introduce a novel structure learning algorithm, and demonstrate its power to accurately reconstruct the regulatory network from those datasets.

Presentation 55

Tuesday, 3 August **1620 - 1650**
Clyde Auditorium

Application of a new probabilistic model for recognizing complex patterns in glycans
Kiyoko F. Aoki, Nobuhisa Ueda, Atsuko Yamaguchi, Minoru Kanehisa, Tatsuya Akutsu and Hiroshi Mamitsuka

Motivation: The study of carbohydrate sugar chains, or glycans, has been one of slow progress mainly due to the difficulty in establishing standard methods for analyzing their structures and biosynthesis. Glycans are generally tree structures that are more complex than linear DNA or protein sequences, and evidence shows that patterns in glycans may be present that spread across siblings and into further regions that are not limited by the edges in the actual tree structure itself. Current models were not able to capture such patterns.

Results: We have applied a new probabilistic model, called probabilistic sibling-dependent tree Markov model (PSTMM), which is able to inherently capture such complex patterns of glycans. Not only is the ability to capture such patterns important in itself, but this also implies that PSTMM is capable of performing multiple tree structure alignments efficiently. We prove through experimentation on actual glycan data that this new model is extremely useful for gaining insight into the hidden, complex patterns of glycans, which are so crucial for the development and functioning of higher level organisms. Furthermore, we also show that this model can be additionally utilized as an innovative approach to multiple tree alignment, which has not been applied to glycan chains before. This extension on the usage of PSTMM may be a major step forward for not only the structural analysis of glycans, but it may consequently prove useful for discovering clues into their function.

Presentation 56

Tuesday, 3 August **1620 - 1650**
Lomond Auditorium

TraitMap: an XML-based genetic-map database combining multigenic loci and biomolecular networks
Naohiko Heida, Yoshikazu Hasegawa, Yoshiki Mochizuki, Katsura Hirosawa, Akihiko Konagaya and Tetsuro Toyoda

Motivation: Most ordinary traits are well described by multiple measurable parameters. Thus, in the course of elucidating the genes responsible for a given trait, it is necessary to conduct and integrate the genetic mapping of each parameter. However, the integration of multiple mapping results from different publications is prevented by the fact that they are conventionally published and accumulated in printed forms or graphics which are difficult for computers to reuse for further analyses.

Results: We have defined an XML-based schema as a container of genetic mapping results, and created a database named TraitMap containing curator-checked data records based on published papers of mapping results in Homo sapiens, Mus musculus, and Arabidopsis thaliana. TraitMap is the first database of mapping charts in genetics, and is integrated in a web-based retrieval framework: termed Genome Phenome Superhighway (GPS) system, where it is possible to combine and visualize multiple mapping records in a two dimensional display. Since most traits are regulated by multiple genes, the system associates every combination of genetic loci to biomolecular networks, and thus helps us to estimate molecular-level candidate networks responsible for a given trait. It is demonstrated that a combined analysis of two diabetes-related traits (susceptibility to insulin resistance and non-HDL cholesterol level) suggests that molecular-level relationships such as the interaction among leptin receptor (Lepr), peroxisome proliferators-activated receptor-gamma (Pparg) and insulin receptor substrate 1 (Irs1), are candidate causal networks affecting the traits in a multigenic manner.

Availability: TraitMap database and GPS are accessible at http://omicspace.riken.jp/gps/

Supplementary information: See http://omicspace.riken.jp/info/traitmap.html

Presentation 57

Tuesday, 3 August **1650 - 1720**
Clyde Auditorium

Centrality of weak interhelical H-bonds in membrane protein functional assembly and conformational gating
Ilan Samish, Eran Goldberg, Oksana Kerner, Avigdor Scherz

Our analysis demonstrates that backbone-mediated interhelical hydrogen-bonds cluster laterally in the conserved core of transmembrane helical proteins. Each residue's propensity to bear these interactions is in correlation with the residue's packing-value scale; giving biophysical meaning to this phenomenological scale. Residues participating in such an intersubunit, structurally conserved H-bond in reaction centers of photosystem II were combinatorially mutated and characterized in silico and in vivo suggesting that H-bond reversible association regulates protein-gated electron transfer. Similar motifs may be involved in folding and conformational flexibility of other membrane proteins. Hence, these findings provide new parameters for structure and function prediction.

Presentation 58

Tuesday, 3 August **1650 - 1720**
Lomond Auditorium

Divergent evolutionary drift contradicts power law
Teresa Przytycka, Yi-Kuo Yu

Recent studies of properties of various biological networks revealed that many of them display scale free characteristics. Since the theory of scale free networks is applicable to evolving networks, one can hope that it provides not only a model of a biological network in its current state but also some insight into the evolution of the network. Here, we re-investigate the probability distributions and scaling properties underlying some models for biological networks and protein domain evolution and point out possible traps in applying a scale free framework to such data. In particular, we demonstrate that divergent evolutionary drift, which is plausible evolutionary mechanisms, is not compatible with scale free models.

Keynote Lecture

Tuesday, 3 August **1720 - 1800**
Clyde Auditorium

Anna Tramontano

Keynote Lecture

Wednesday, 4 August **0830 - 0920**
Clyde Auditorium

Overton Prize Lecture
Uri Alon

Presentation 59

Wednesday, 4 August **0920 - 0940**
Clyde Auditorium

Constraint-based modelling of perturbed organisms: a ROOM for improvement
Tomer Shlomi, Omer Berkman, Eytan Ruppin

Regulatory On-Off Minimization (ROOM) is a model for predicting the behavior of metabolic networks in response to gene knockouts. It is based on minimizing the number of significant flux changes (hence, on/off) with respect to the wild-type. ROOM outperforms a previously suggested model, Minimization Of Metabolic Adjustment (MOMA), whose minimization metric is Euclidian. Furthermore, ROOM shows its ability to correctly identify alternative pathways for reactions associated with the knocked-out genes, thus strengthening its biological plausibility. ROOM outperforms MOMA in predicting intracellular fluxes and gene knockout lethality in mutated E. coli and the S. cerevisiae strains, respectively.

Presentation 60

Wednesday, 4 August **0940 - 1010**
Clyde Auditorium

Filling gaps in a metabolic network using expression information
Peter Kharchenko, Dennis Vitkup and George M. Church

Motivation: The metabolic models of both newly sequenced and well-studied organisms contain reactions for which the enzymes have not been identified yet. We present a computational approach for identifying genes encoding such missing metabolic enzymes in a partially reconstructed metabolic network.

Results: The metabolic expression placement (MEP) method relies on the coexpression properties of the metabolic network and is complementary to the sequence homology and genome context methods that are currently being used to identify missing metabolic genes. The MEP algorithm predicts over 20% of all known Saccharomyces cerevisiae metabolic enzyme-encoding genes within the top 50 out of 5594 candidates for their enzymatic function, and 70% of metabolic genes whose expression level has been significantly perturbed across the conditions of the expression dataset used.

Availability: Freely available (in Supplementary information).

Supplementary information: Available at the following URL
http://arep.med.harvard.edu/kharchenko/mep/supplements.html

Presentation 61

Wednesday, 4 August **0940 - 1010**
Lomond Auditorium

GeneXPress: a visualization and statistical analysis tool for gene expression and sequence data
Eran Segal, Amit Kaushal, Roman Yelensky, Tuan Pham, Aviv Regev, Daphne Koller, Nir Friedman

Many algorithms have been developed for analyzing gene expression and sequence data. However, to extract biological understanding, scientists often have to perform further time consuming post-processing on the output of these algorithms. In this paper, we present GeneXPress, a tool designed to facilitate the assginment of biological meaning to gene expression patterns by automating this post processing stage. Within a few simple steps that take at most several minutes, a user of GeneXPress can: identify the biological processes represented by each cluster; identify the DNA binding sites that are unique to the genes in each cluster; and examine multiple visualizations of the expression and sequence data. GeneXPress thus allows the researcher to quickly identify potentially new biological discoveries. GeneXPress is available for download at http://GeneXPress.stanford.edu.

Presentation 62

Wednesday, 4 August **1010 - 1040**
Clyde Auditorium

An efficient algorithm for detecting frequent subgraphs in biological networks
Mehmet Koyutürk, Ananth Grama and Wojciech Szpankowski

Motivation: With rapidly increasing amount of network and interaction data in molecular biology, the problem of effectively analyzing this data is an important one. Graph theoretic formalisms, commonly used for these analysis tasks, often lead to computationally hard problems due to their relation with subgraph isomorphism.

Results: This paper presents an innovative new algorithm for detecting frequently occurring patterns and modules in biological networks. Using an innovative graph simplification technique, which is ideally suited to biological networks, our algorithm renders these problems computationally tractable. Indeed, we show experimentally that our algorithm can extract frequently occurring patterns in metabolic pathways extracted from the KEGG database within seconds. The proposed model and algorithm are applicable to a variety of biological networks either directly or with minor modifications.

Availability: Implementation of the proposed algorithms in the C programming language is available as open source at http://www.cs.purdue.edu/homes/koyuturk/pathway/

Presentation 63

Wednesday, 4 August **1010 - 1040**
Lomond Auditorium

Filtering erroneous protein annotation
D. Wieser, E. Kretschmann and R. Apweiler*

Motivation: Automatically generated annotation on protein data of UniProt (Universal Protein Resource) is planned to be publicly available on the UniProt web pages in April 2004. It is expected that the data content of over 500 000 protein entries in the TrEMBL section will be enhanced by the output of an automated annotation pipeline. However, a part of the automatically added data will be erroneous, as are parts of the information coming from other sources. We present a post-processing system called Xanthippe that is based on a simple exclusion mechanism and a decision tree approach using the C4.5 data-mining algorithm.

Results: It is shown that Xanthippe detects and flags a large part of the annotation errors and considerably increases the reliability of both automatically generated data and annotation from other sources. As a cross-validation to Swiss-Prot shows, errors in protein descriptions, comments and keywords are successfully filtered out. Xanthippe is a contradictive application that can be combined seamlessly with predictive systems. It can be used either to improve the precision of automated annotation at a constant level of recall or increase the recall at a constant level of precision.

Availability: The application of the Xanthippe rules can be browsed at http://www.ebi.uniprot.org/

Presentation 64

Wednesday, 4 August **1110 - 1140**
Clyde Auditorium

A knowledge based approach for representing and reasoning about signaling networks
C. Baral, K. Chancellor, N. Tran, N.L. Tran, A. Joy and M. Berens

Motivation: In this paper we propose to use recent developments in knowledge representation languages and reasoning methodologies for representing and reasoning about signaling networks. Our approach is different from most other qualitative systems biology approaches in that it is based on reasoning (or inferencing) rather than simulation. Some of the advantages of our approach are, we can use recent advances in reasoning with incomplete and partial information to deal with gaps in signal network knowledge; and can perform various kinds of reasoning such as planning, hypothetical reasoning and explaining observations.

Results: Using our approach we have developed the system BioSigNet-RR for representation and reasoning about signaling networks. We use a NFkB related signaling pathway to illustrate the kinds of reasoning and representation that our system can currently do.

Availability: The system is available on the Web at http://www.public.asu.edu/~cbaral/biosignet

Presentation 65

Wednesday, 4 August **1110 - 1140**
Lomond Auditorium

Selecting biomedical data sources according to user preferences
Sarah Cohen Boulakia, Séverine Lair, Nicolas Stransky, Stéphane Graziani, François Radvanyi, Emmanuel Barillot and Christine Froidevaux

Motivation: Biologists are now faced with the problem of integrating information from multiple heterogeneous public sources with their own experimental data contained in individual sources. The selection of the sources to be considered is thus critically important.

Results: Our aim is to support biologists by developing a module based on an algorithm that presents a selection of sources relevant to their query and matched to their own preferences. We approached this task by investigating the characteristics of biomedical data and introducing several preference criteria useful for bioinformaticians. This work was carried out in the framework of a project which aims to develop an integrative platform for the multiple parametric analysis of cancer. We illustrate our study through an elementary biomedical query occurring in a CGH analysis scenario.

Availability: http://www.lri.fr/~cohen/dss/dss.html

Presentation 66

Wednesday, 4 August **1140 - 1210**
Lomond Auditorium

Integration of biological data from web resources: management of multiple answers through metadata retrieval
Marie-Dominique Devignes, Malika Smail

Biological data retrieval from web resources often necessitates multi-step access to multiple information sources. User-designed scenarios are exploited by a generic application (Xcollect) that allows users to execute them and to store the collected data in a document. Multiple answers are obtained at given steps of the scenario when several resources are queried with same purpose. We address the problem of managing such multiple answers when retrieving functional annotations of genes. Relevant quality metadata for sources and source entries have been listed in view of sorting the answers. Work in progress deals with semantic integration based on domain ontologies.

Presentation 67

Wednesday, 4 August **1210 - 1230**
Lomond Auditorium

Criticality-based task composition in distributed bioinformatics systems
Konstantinos Karasavvas, Richard Baldock, Albert Burger

During task composition, such as can be found in distributed query processing, workflow systems and AI planning, decisions have to be made by the system and possibly by users with respect to how a given problem should be solved. Although there is often more than one correct way of solving a given problem, these multiple solutions do not necessarily lead to the same result. Some researchers are addressing this problem by providing data provenance information. In this paper we propose an approach that assesses the importance of such decisions with respect to the overall result. We present a way of measuring decision criticality and describe its potential use. Real bioinformatics examples are used to illustrate the approach.

Keynote Lecture

Wednesday, 4 August **1400 - 1450**
Clyde Auditorium

ISCB Senior Scientist Accomplishment Award Lecture
David Lipman