Presented by Xue-wen Chen

Director of BCLSL, ITTC

ITTC IAB, June 10, 2010

Research Overview
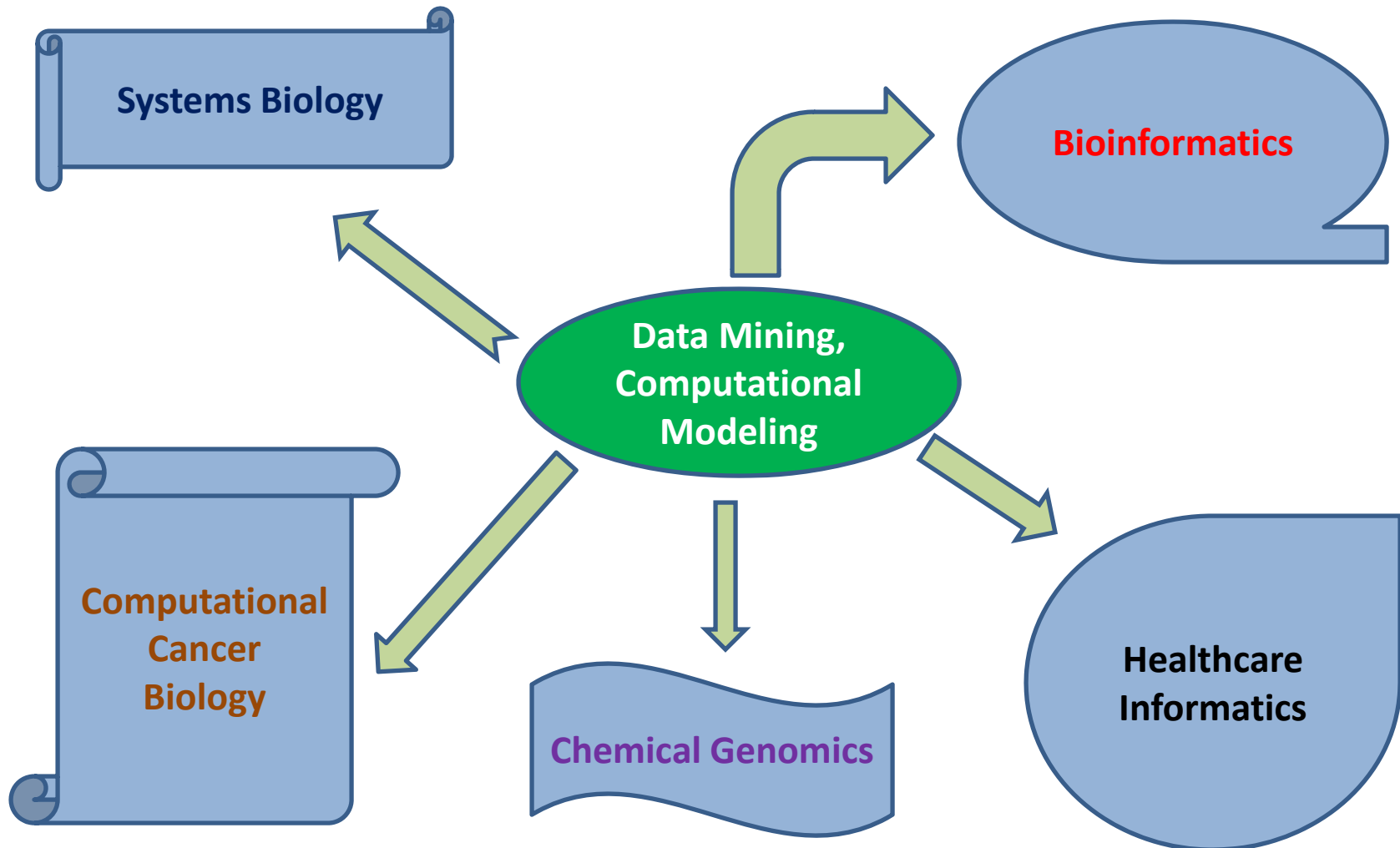
Bioinformatics

- "The mathematical, statistical and computing methods that aim to solve biological problems using DNA and amino acid sequences and related information."

- Example: Protein function prediction
- Example: Binding site prediction

I. Initial DDI Network Construction

Cross-species PPI network

Cluster PPIs into functionally "similar" groups

....

Infer the top scoring DDI from each group

....

Assemble statistically significant DDIs into a network

Find partial-complete bipartite graph with the biclustering algorithm, and fill-in the missing DDIs

II. DDI Network Expansion

Compile a set of proteins sharing same GO term and their interaction partners

Extract domains

Extract network spanning these domains

Example: predicting mutations enhancing protein stability

The problem: native proteins are usually only marginally stable at ambient temperature; consequently shelf lives of protein pharmaceuticals are often not long enough, enzymes cannot work at elevated temperatures...

Possible Solutions:

- Adding protein stabilizers

- Engineering proteins: random screening, directed evolution, **rational design**

 Our approach: rational design using novel predictive models based on large scale protein structure and sequence mining

Possible applications: pharmaceuticals, bio-energy, oil and gas, paper and pulp, many more.

Systems Biology

- Genes, proteins, cells, tissues and organisms form complex networks of interacting systems to function
- Traditionally, study of individual biological and biochemical processes generates a wealth of knowledge
  - Networks are often constructed on conjecture
  - Relatively limited insights about the biological system as a whole
- Recent advances in high-throughput technologies (e.g., genome sequencing, Y2H, microarray, …)
  - Data rich, information rich
  - Systems-level research
  - Left with more questions than answers
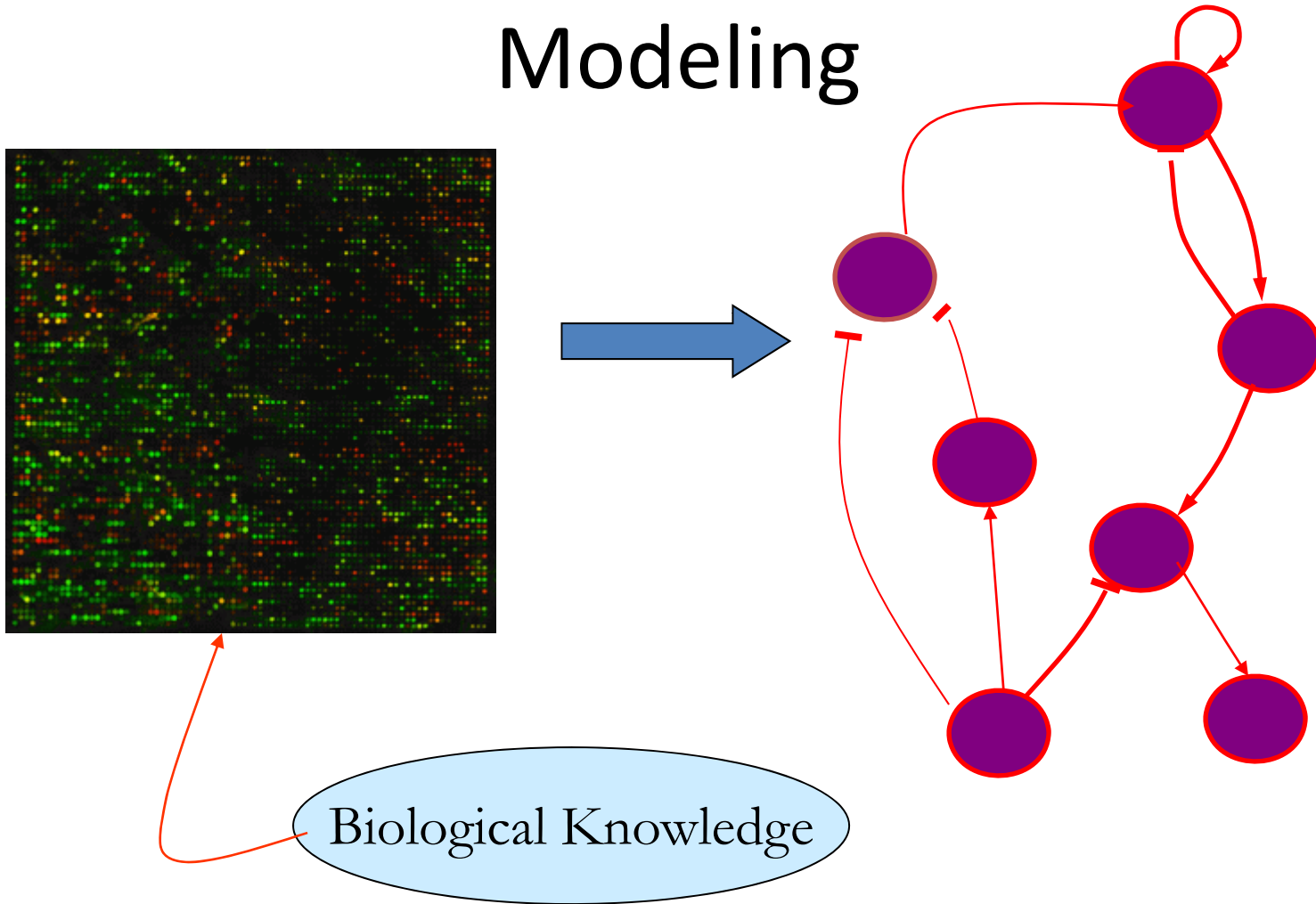
Computational Systems Biology

- CSB focuses on developing computational models for the systematic study of complex interactions in biological systems (networks)
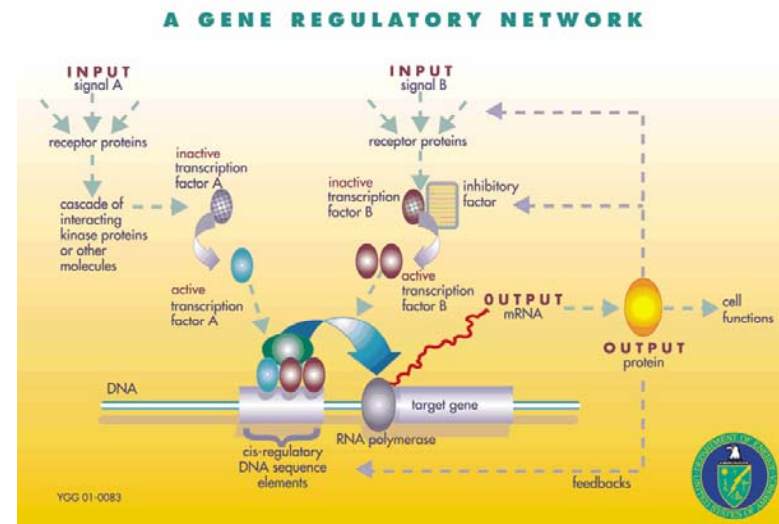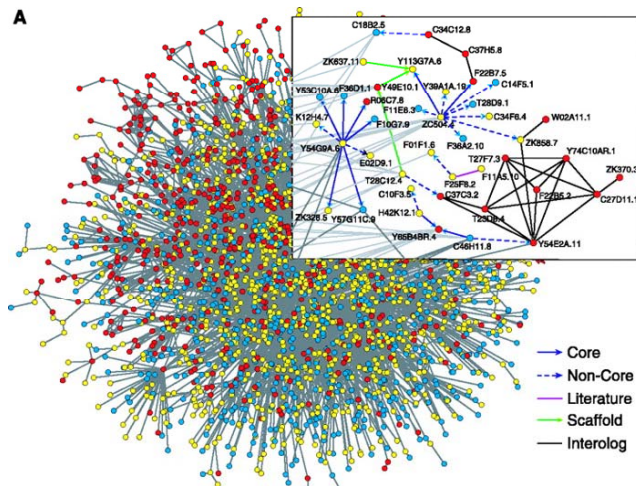
C. elegans: Li et al., Science, 303: 540-543, 2004
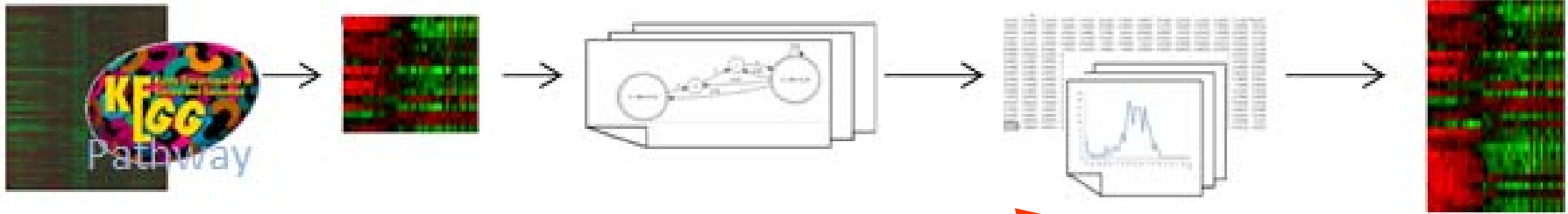

US DOE Genomics: GTL Program

Systems Biology: Computational Modeling
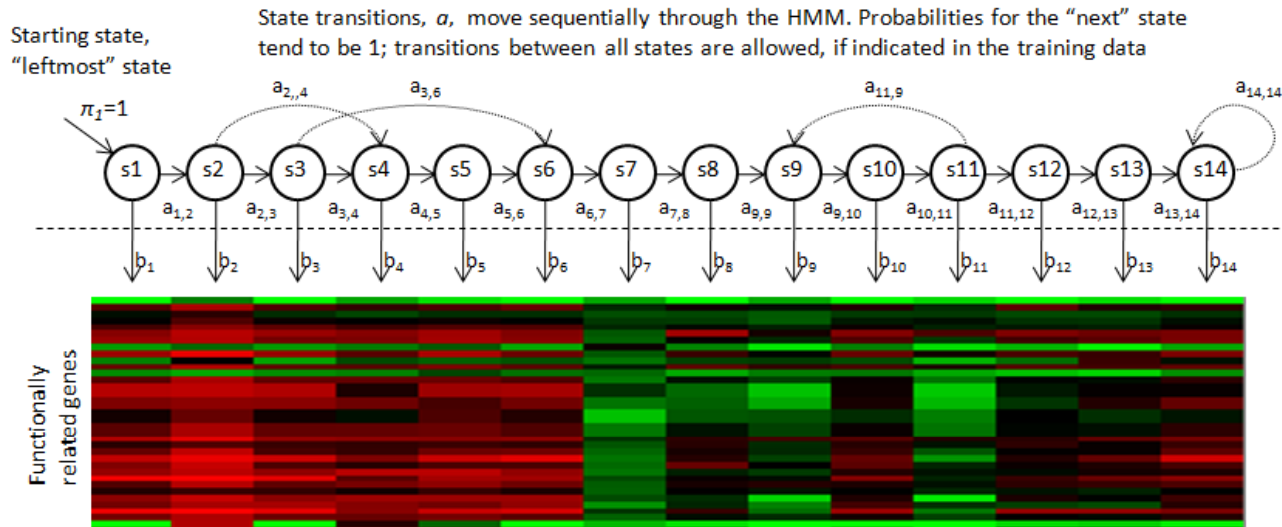

Biological Knowledge

Identifying Genes in the Same Functional Module

Select conditions   Multiple HMM

the Parzen density distribution function (PDF) $\hat{p}(x) = \frac{1}{V} \sum_{i=1}^{n} K_i(x_i)$

Starting state, "leftmost" state

State transitions, *a*, move sequentially through the HMM. Probabilities for the "next" state tend to be 1; transitions between all states are allowed, if indicated in the training data

Emission probabilities, *b*, are the probability distributions for gene expression values at each column in the feature set
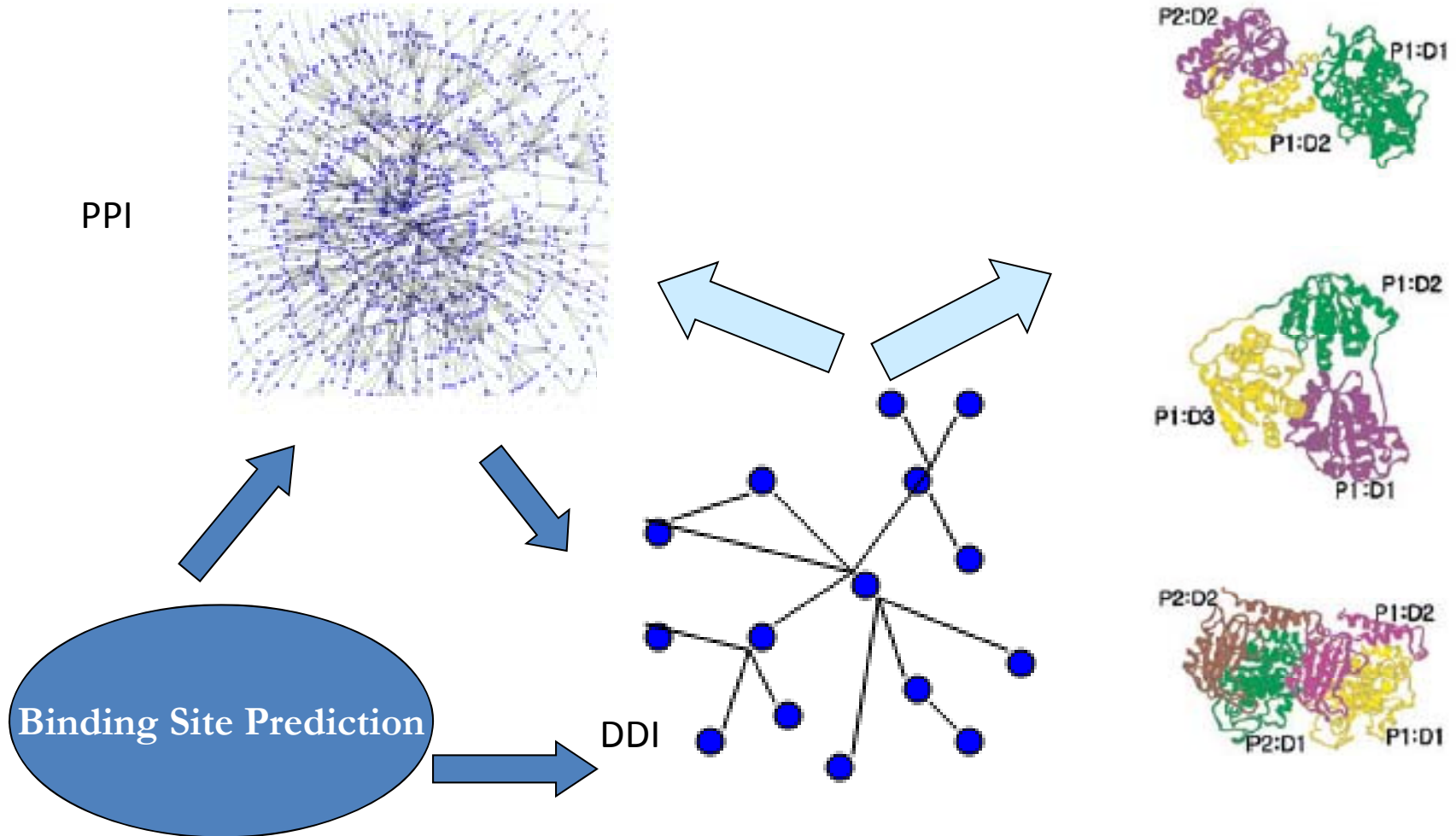
BN-based Network Reconstruction

- $P(x_1, x_2, x_3, x_4, x_5) = P(x_1)P(x_2|x_1)P(x_3|x_1)P(x_4|x_3)P(x_5|x_1, x_2, x_3)$
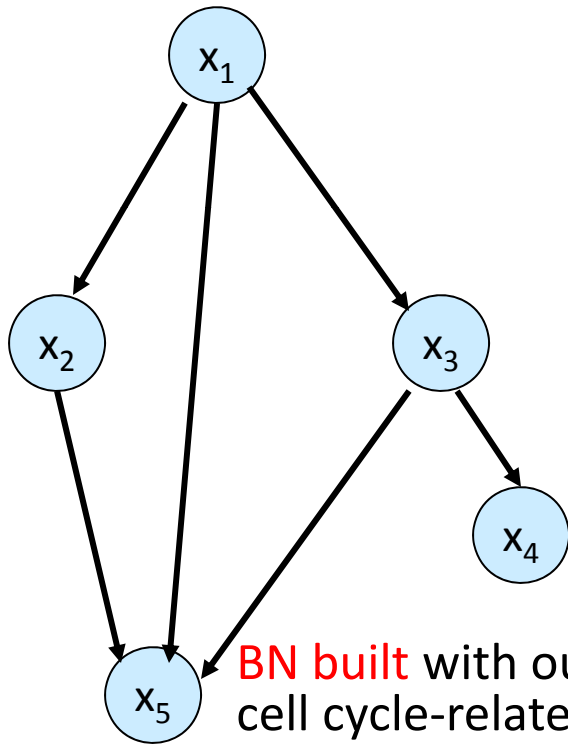
BN built with our algorithm for yeast cell cycle-related genes. A total of 20 nodes (genes) and 34 edges (interactions) were incorporated. This network captured 65% of all currently reported direct and indirect interactions among these genes.
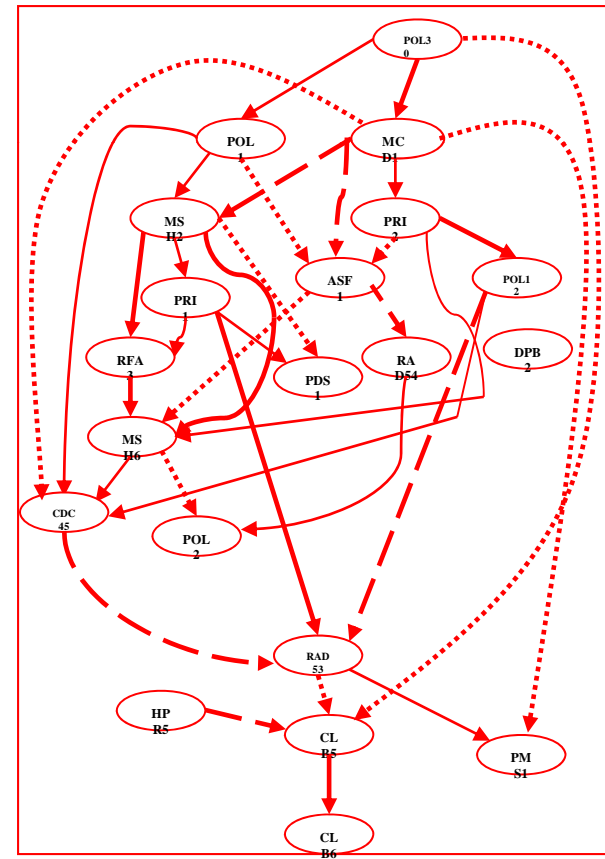
Domain/Protein Interaction Networks

Computational Cancer Biology

Serum Proteins

Protein chip

SELDI-TOF MS

m/z
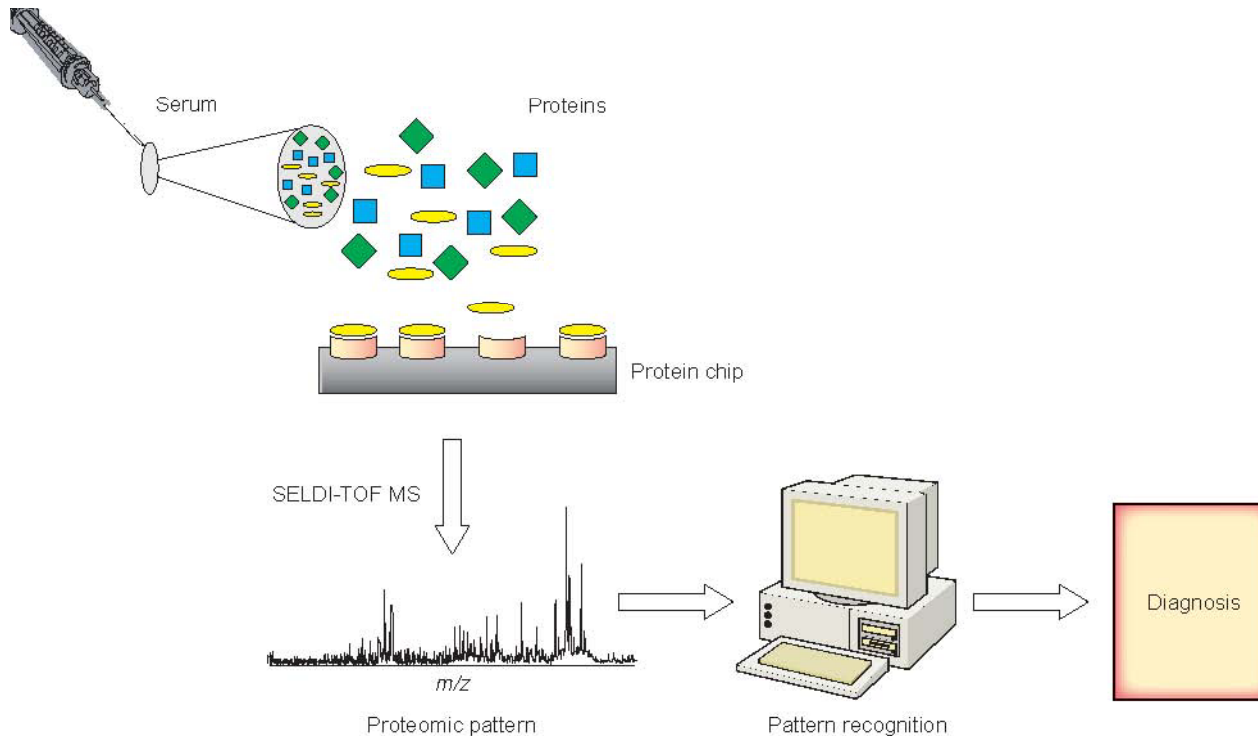
Proteomic pattern Pattern recognition Diagnosis

Figure 1. Disease diagnostics using proteomic patterns. The sample drawn from the patient is applied to a protein chip which is made up of a specific chromatographic surface. After several washing steps and the application of an energy-absorbing molecule, the species that are retained on the surface of the chip are analyzed via mass spectrometry. The pattern of peaks within the spectrum is analyzed using sophisticated bioinformatic software to diagnose the source of the biological sample.

m/z: Mass to charge ratio; SELDI-TOF MS: Surface-enhanced laser desorption/ionization time-of-flight mass spectrometry.

- Example: disease diagnosis using proteomic patterns

Example: Computer Assisted Cancer Diagnosis

- Visual features in tissue biopsies are heavily relied on for cancer diagnosis.

- Hours examining mostly healthy tissue can result in fatigue and human error.

- For some forms of cancer, expert diagnosis from a set of biopsies may vary widely between experts.

- Automated methods can prune obviously healthy tissue from human search, and reduce variability among experts.

Different grades of prostate cancer are distinguished by clear visual features. However, many individual cases are ambiguous.

Computer Assisted Cancer Diagnosis

- Existing methods of computer-assisted cancer diagnosis rely on ad hoc, manually tuned visual features.

- Our work focuses on the automated extraction of visual features that are optimally discriminative towards cancer types and grades.

- We are working to develop hierarchical probabilistic graphical models capable of extracting complex higher-order statistical relationships that can be used to more effectively and more reliably diagnose cancer.

Different grades of prostate cancer are distinguished by clear visual features. However, many individual cases are ambiguous.

Other "Disease"-related Research

- Computational identification of disease genes
  - (e.g., using PPIs, microarray, etc.)
- Genome-wide Association Study (GWAS)
 - (e.g., genotyped data analysis, SNP-disease association in autism etc.)
- Pathway analysis
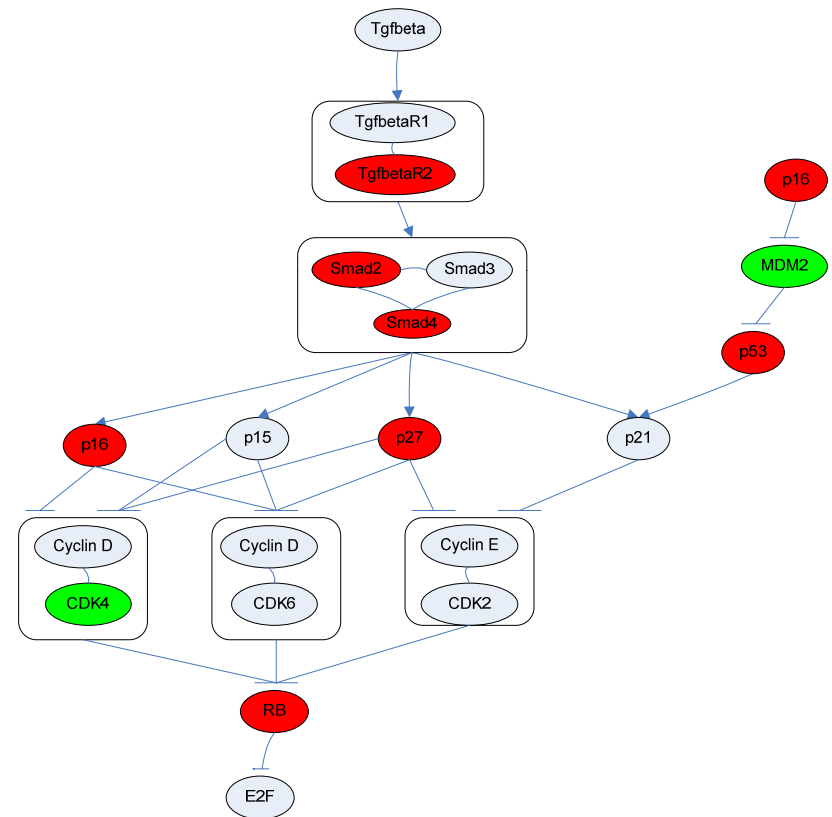
Antigrowth signaling pathway

Chemical Genomics

- Chemical genomics studies the interaction of small molecules with cellular systems *in vitro* and *in vivo*.
- Goal: developing advanced informatics approach supporting the management, analysis, data mining, and transformation of experimental data to predictive models and to identify biologically active compounds that may modulate biological systems.
- Focusing on *prediction* and *reconstruction* of chemical interaction networks
  - Protein-chemical interaction prediction
 - Gene-chemical interaction prediction

Graph based QSAR-Tox Study

- Use graph representation rather than numeric properties.
- Transformation of chemicals to graphs is straight forward.
  - Atoms correspond to vertices.
  - Bonds correspond to edges.
  - Vertices and edges are labeled with atom element and bond type, among other properties.

Chemical Graphs
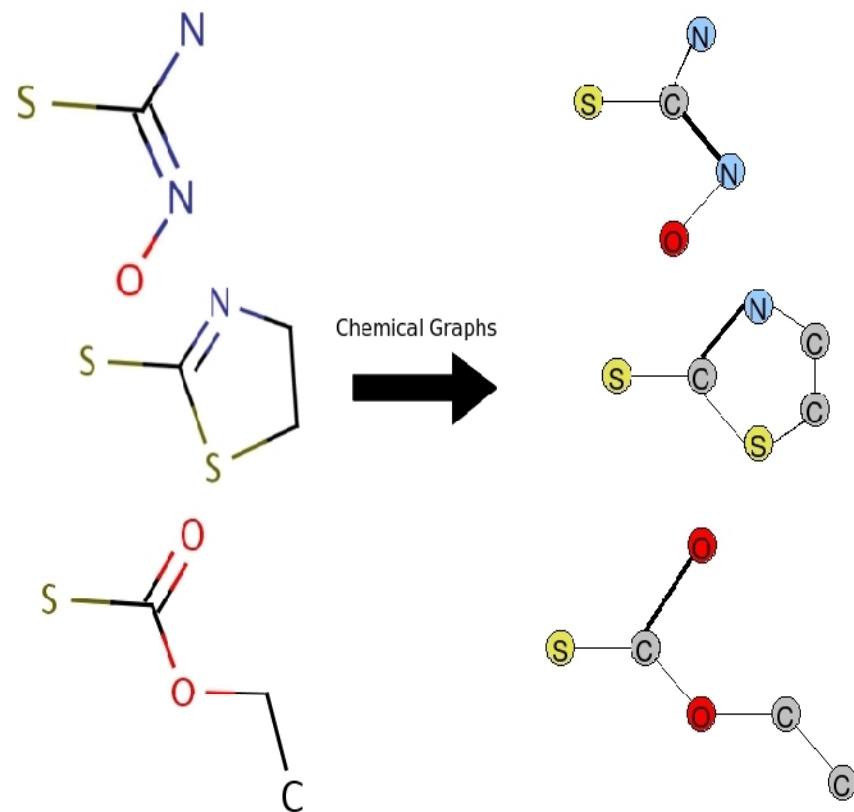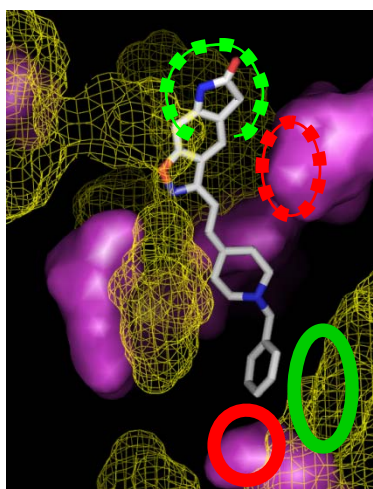
Learning with Graphs

- Node labeling with membership test
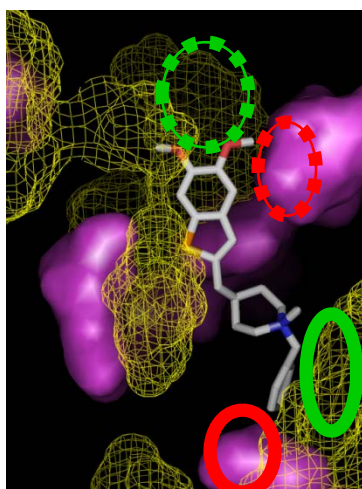 - Each node in a graph is labeled with a vector of bits indicating *memberships* to a set of features
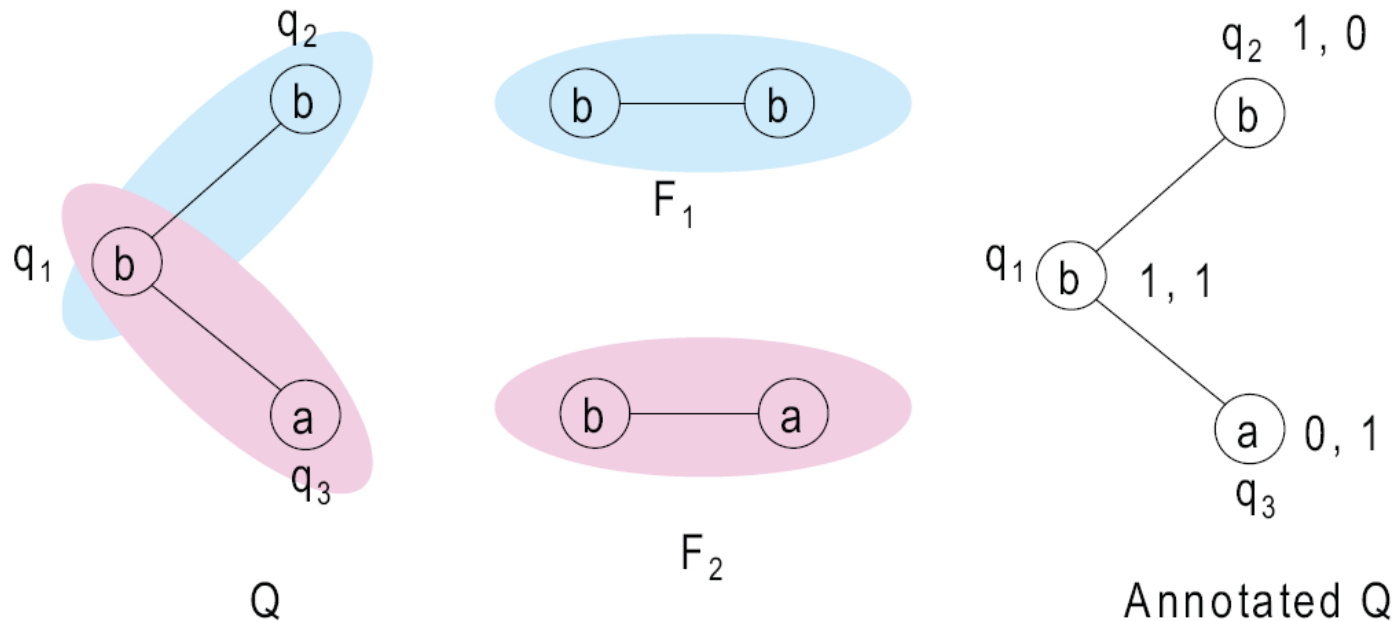
Computational Pharmacology

Structure-based Drug Design
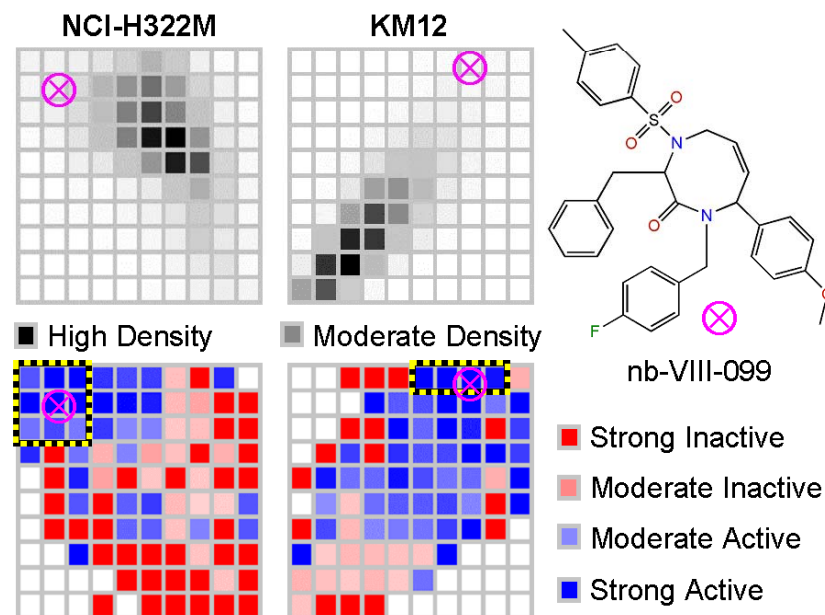
Potent   Weak

Train predictive pharmacological models to reproduce chemical biology screening data using descriptors based on properties of the enzyme-inhibitor complex

Chemical Diversity Analysis

NCI-H322M   KM12

■ High Density   ■ Moderate Density

nb-VIII-099

■ Strong Inactive
■ Moderate Inactive
■ Moderate Active
■ Strong Active

Guide chemical synthesis efforts via Cartesian chemical property space models that distinguish compounds according to chemical novelty and biological relevance

Healthcare Informatics

- New field
- Web-enabled technology (mining, knowledge discovery)
- Health and wellness informatics
- Computational Support for evidence-based care
- Evaluation of health information system
- EMR-based applications (e.g., keyword-based search, knowledge discovery)
- Mining of medical records

Cephalometric Benchmark Tracing

❑Cephalometric landmarks identification and tracing is very important to orthognathic surgery.

❑Identifying the medical points, lines, and faces is a difficult and boring task for doctors.

—X-ray images are not always clearly projected.

—Low accuracy in manual identification and plotting.
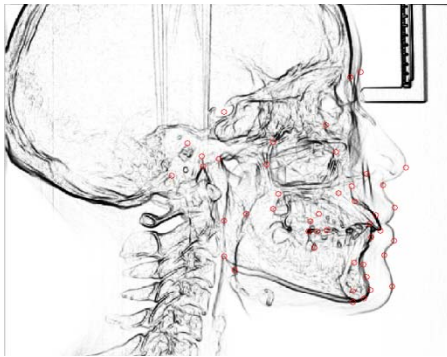
—Labor intensive – hundreds of cases per doctor per month.

Sample points which are manually marked

Table 1: Landmarks for Cephalometric Tracing

| Cranial	Supramaxilla	Mandible	Soft tissue
N. Nasion S. Sella Ba. Basion Bolton	O. Orbitale ANS. Anterior Nasal Spine PNS. Posterior Nasal Spine Ptm. Pterygomaxillary Fissure A. Subspinale SPr. Superior Prosthion UI. Upper Incisor	Co. Condylion Ar. Articulare Go. Gonion B. Supramentale Id. Infradentale Li. Lower Incisor Po. Pogonion Me. Menton Gn. Gnathion	G. Glabella Ns. Nasion of Soft Tissue E. Eye Prn Sn. Subnasale UL LL Pos. Pogonion of Soft Tissue

Cephalometric Benchmark Tracing

❑ Canny edge detector(soft & hard contour), Hough transform (Sella circle, upper jaw), etc.

   ❑ Soft tissue detection using deformable template

   ❑ Automatic localization of landmarks for hard tissue

❑ Will build an integrative system for analyzing both images and text
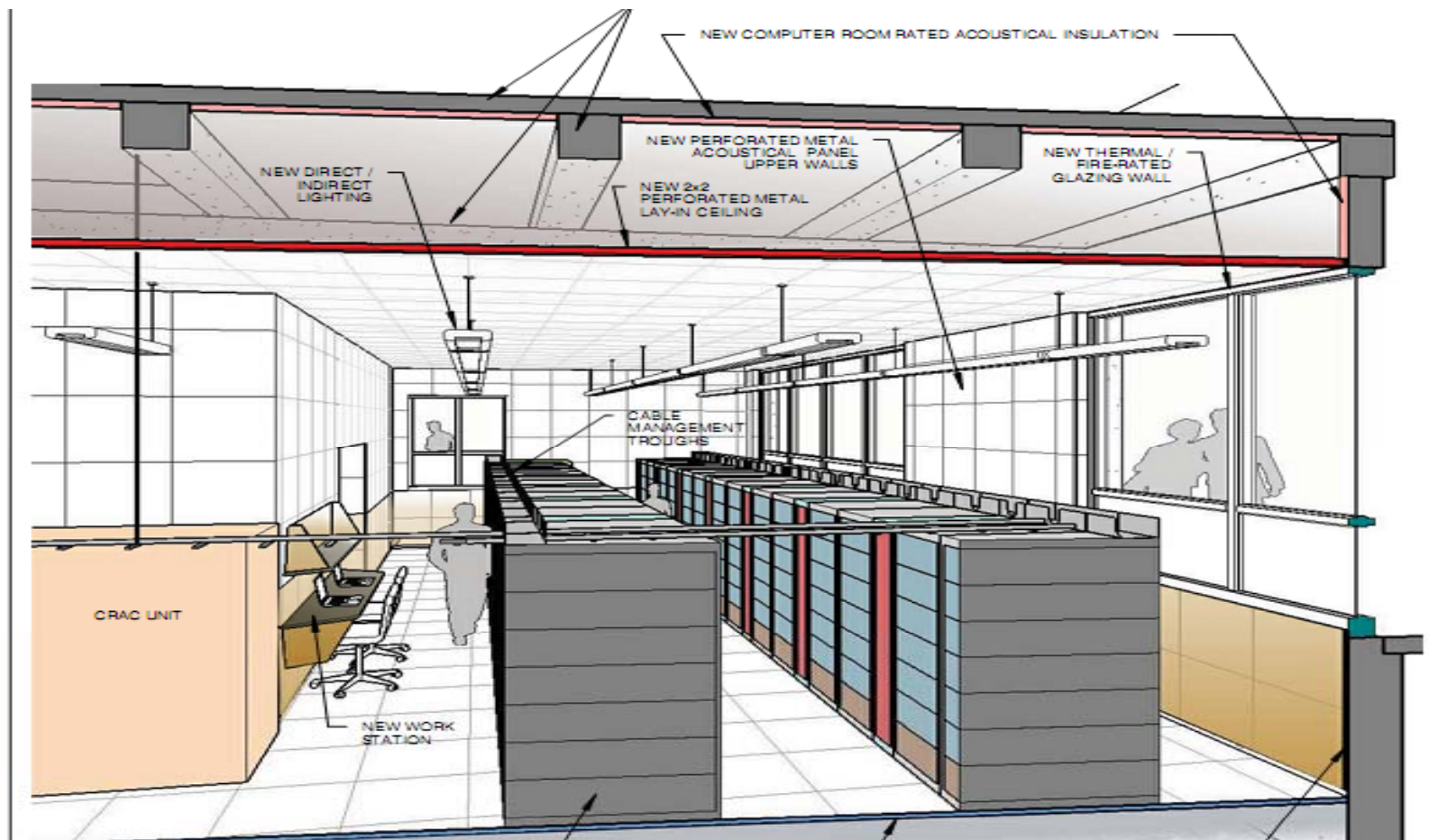
Sample of cephalometric sketch

Soft tissue deteiction using deformable template
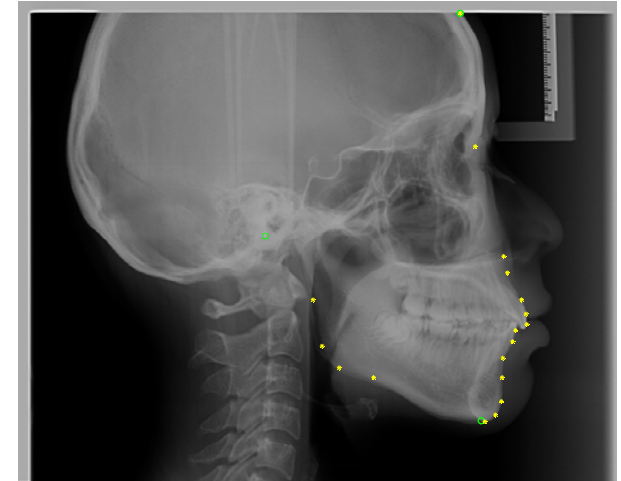
Hard tissue detection using deformable template

Next-Generation High Performance Computing Facility

KU received a $4.8 M NIH G20 fund to upgrade the cold room of the existing Bioinformatics Computing Facility (BCF). Luke Huan plays a significant role in leading the KU G20 proposal development

Acknowledgment

- Some of the slides are provided by
  - Jianwen Fang
  - Luke Huan
  - Bo Luo
  - Gerry Lushington
  - Brian Potetz