Inferring Networks from Distances: the "Landscape" of Glycosidase Protein Structures

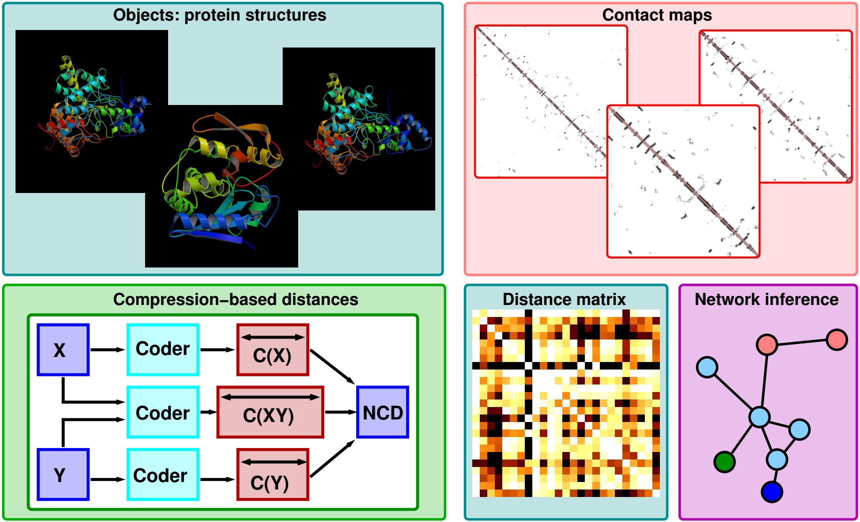
Sandhya Prabhakaran, Volker Roth

Department of Mathematics and Computer Science,
University of Basel, Switzerland,

UNI
BASEL

Nov 21, 2011

Graphical Abstract



Objects: protein structures

Contact maps

Compression–based distances

X

Y

Coder

Coder

Coder

C(X)

C(XY)

C(Y)

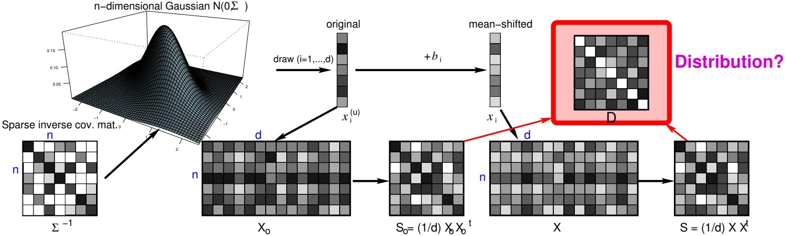
NCD

Distance matrix

Network inference

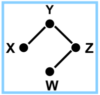
Bayesian Inference

- Idea: construct **probabilistic generative model**, **condition on observed data** and **infer distribution of parameters.**



- **We observe pairwise distances** D.
- $D \sim$ **singular Wishart**, parametrized by inverse covariance Ψ.
- **Link to Gaussian graphical models:**

  $\Psi_{ij} = 0 \rightsquigarrow (i,j)$ **conditionally independent**
  $\rightsquigarrow$ **no edge**.



- Infer Ψ by **Markov Chain Monte Carlo sampling**.

Compression Distance between Protein Structures

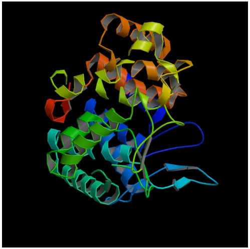
- **Normalized Information Distance**: length of the shortest program that computes $x|y$ and $y|x$.

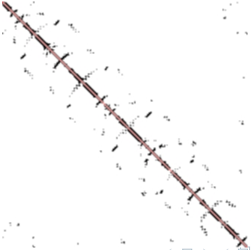
$$NID(x,y) \propto \max\{K(x|y), K(y|x)\} = K(xy) - \min\{K(x), K(y)\}$$

- Approximation: **Normalized Compression Distance:**

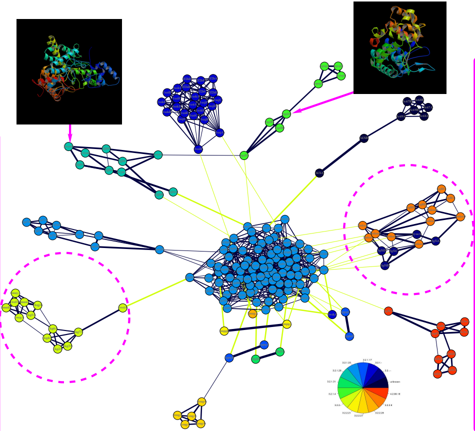
$$NCD(x,y) \propto C(xy) - \min\{C(x), C(y)\}.$$

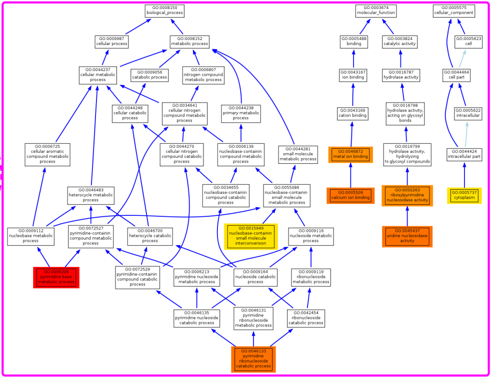
- **In our application:** x, y are vectorized **contact maps** for **Glycosidase enzymes** in **Escherichia coli**.





The “Landscape” of Glycosidase Protein Structures





GO:0008150 biological_process

GO:0009987 cellular process

GO:0008152 metabolic process

GO:0044237 cellular metabolic process

GO:0009056 catabolic process

GO:0006807 nitrogen compound metabolic process

GO:0044248 cellular catabolic process

GO:0034641 cellular nitrogen compound metabolic process

GO:0044238 primary metabolic process

GO:0006725 cellular aromatic compound metabolic process

GO:0044270 cellular nitrogen compound catabolic process

GO:0006139 nucleobase-containin compound metabolic process

GO:0044281 small molecule metabolic process

GO:0046483 heterocycle metabolic process

GO:0034655 nucleobase-containin compound catabolic process

GO:0055086 nucleobase-containin small molecule metabolic process

GO:0009112 nucleobase metabolic process

GO:0072527 pyrimidine-containin compound metabolic process

GO:0046700 heterocycle catabolic process

GO:0015949 nucleobase-containin small molecule interconversion

GO:0009116 nucleoside metabolic process

GO:0006206 pyrimidine base metabolic process

GO:0072529 pyrimidine-containin compound catabolic process

GO:0006213 pyrimidine nucleoside metabolic process

GO:0009164 nucleoside catabolic process

GO:0009119 ribonucleoside metabolic process

GO:0046135 pyrimidine nucleoside catabolic process

GO:0046131 pyrimidine ribonucleoside metabolic process

GO:0042454 ribonucleoside catabolic process

GO:0046133 pyrimidine ribonucleoside catabolic process

GO:0003674 molecular_function

GO:0005488 binding

GO:0003824 catalytic activity

GO:0043167 ion binding

GO:0016787 hydrolase activity

GO:0043169 cation binding

GO:0016798 hydrolase activity, acting on glycosyl bonds

GO:0046872 metal ion binding

GO:0016799 hydrolase activity, hydrolyzing N-glycosyl compounds

GO:0005509 calcium ion binding

GO:0050263 ribosylpyrimidine nucleosidase activity

GO:0045437 uridine nucleosidase activity

GO:0005575 cellular_component

GO:0005623 cell

GO:0044464 cell part

GO:0005622 intracellular

GO:0044424 intracellular part

GO:0005737 cytoplasm