Automated prediction of protein function and detection of functional sites from structure

By Florencio Pazos and Michael Sternberg

Junguk Hur
School of Informatics
Background-I

- Increasing portion of protein structure without knowledge of function.
  - “Hypothetical” ➔ 1,142 / 35,917 ≈ 3.2%
    (Apr. 5, 2006)
  - Will increase more due to structural genomics

- Current structure – function gap
  ➔ More powerful bioinformatics techniques urgently needed

Junguk Hur
Automated function prediction
**Background-II**

- **Sequence based function prediction**
  - Identification of a high sequence similarity
  - Specific profiles and related hidden Markov models
  - Subfamily-specific functional sites

- **Structural information**
  - Insight into protein function
  - Similar 3D folds without sequence identity
    - May not guarantee similar function
    - Must identify sequence and spatial location of key functional residues

Junguk Hur  
Automated function prediction
Background-III

- Use of 3D templates associated to function.
- Identification of functional sites and regions
  - Fully conserved residues
  - Family specific conservation
  - 3D structural information alone
  - 3D structural information with sequence information
- A recent method (graph-theoretic) for automatic identification of 3D profiles without structural alignment (Wangikar et al, 2003)
  ➔ successful but still limited
Automated function prediction method via identification of functional sites

Based on several concepts
- More accurate using structural alignment info
- GO classification to group similar proteins
- GO hierarchical and extensive classification
- Residue responsible for specificity of a subfamily will only be conserved within that subfamily
GO (Gene Ontology)

- Gene ontology consortium (http://www.geneontology.org/)
- Hierarchically well structured vocabulary of biological terms
PHUNCTIONER

Junguk Hur

Automated function prediction
Function determining residues

- FSSP (Oct. 2001) preprocessing (removal)
  - $\geq 35\%$ SID
  - Structural similarity $< 6.0$ (FSSP Z score)
  - Annotated as ‘mutant’
- Subalignments; one for each GO term
  - Conservation
    $$C_i = \frac{\text{sim}(a_{ij}, a_{ik})}{\text{McLachlan substitution matrix}}$$
    (spatial structure)
  - Z-score to assess how conserved in whole subalignment
    $$Z_i = \frac{C_i - \bar{C}}{\sigma},$$
  - Thresholds: $< Z_{\text{cut}}$ OR $> Z_{\text{ocut}}$

$\Rightarrow$ “functional determinant” of a given GO “function”
Function determining residues

- Construction of PSSM

\[ P_{ij} = \sum_{k=1}^{20} W_{ki} \cdot \text{sim}(k, j), \]

\[ W_{ki} = \frac{\ln \left[ 1 - \left( \frac{f_{ki}}{N + 1} \right) \right]}{\ln \left[ \frac{1}{N + 1} \right]}, \]

- Score of a set of n-residue against PSSM

\[ S = \sum_{i=1}^{n} P_{ik}, \quad \Rightarrow \text{5000 iteration} \]

\[ Z = \frac{S - \bar{S}}{\sigma_s}. \]
Testing the Method

- For each sequence in a given filtered FSSP structural alignment
  - Step 1. One sequence is removed
  - Step 2. Rebuilt the subalignments and PSSMs
  - Step 3. Calculate Z-score of the left-aside sequence against each one of the PSSMs

- Similar process using sequence identity.

- To compare two methods
  - At least one profile
  - At least one sequence in the lists
Accurate if the first hit is correct

PHUNCTIONER outperforms SEQID ➔ Good for low SID
• Estimation of true- and false-positive ratio

Junguk Hur

Automated function prediction
Z-score correlation
Examples of functional assignment

- Hypothetical protein MTH538 (PDB:1EIW)
- Has structures of
  - Two-component system receiver domains (CheY)
  - Nucleotide triphosphate (NTP) binding proteins
  - Flavodoxins
- Known fact: binds to Mg$^{2+}$ but not flavin mononucleotide
- **PSI-BLAST** → COG1618 Putative ATPase or kinases
  - But MTH538 has no classical NTP binding domain in MTH538
Examples of functional assignment

- **PHUNCTIONER**
  - “two-component response regulatory activity” (Z=4.96)
  - “Mg2+ ion binding activity” (Z=7.07)
  - No profile regarding “flavin”
  - **Rejects** NTP binding functions (Z=0.8 ~ 1.71)

  → Clearly detects that the protein **does not** fit in the NTP binding motifs, even with global signal.

- **Characteristics**
  - Fully automatic
  - Additional finding
    - Ca2+ binding (Z=5.18)
    - Transcription factor (Z=3.44), transferase (Z=2.10) – False Positive
Extracted Function-Determinant Residues

- Detection of GO associated residues
  - “GTP-binding activity”
  - *Ras* oncogene (PDB: 1CTQ)

Very specific profile
(average Z=7.1)
Comparison with other tools

- Conserved residues of *Ras* family
  - Alignments from HSSP database
  - Redundancy removed (>95%)
  - Three cutoffs of conservation by HSSP VAR
Comparison - continued

- Family dependant conserved residues
  - SEQUENCESPACE (Casari et al.)
  - MTREEDEDET (del sol mesa et al.)

Unable to extract residues exclusively responsible for a GO function

Junguk Hur
Automated function prediction
Limitations

- Can’t generate a profile for a function on more than one fold by convergent evolution
- Requirement of structural information with at least four homologue
- Limited to functions of proteins with known structure
Summary

- Automatic function prediction of structurally known proteins by detecting functional sites

- Automatic functional residue detection for GO terms (functions)

- Especially useful in the twilight zone of function annotation
Thanks