Annotation & Inference
New genomes, New functions

**Having Function**
- Experiments
- Literature
- Expert view

**Maybe**
- Boarder line similarity
- Only part of protein
- Conflicting exp/lit

**Wrong**
- Fault annotation
- Wrong inference

**No Function**
- New genomes
- No similarity
- No evidence

Michal Linial, Institute of Life Sciences
The Hebrew University of Jerusalem

The Hebrew University of Jerusalem

May 2006
Annotation & Inference
New genomes, New functions

Domain families by EVEREST
Automatic identification of Protein Domain
Performance and analysis w.r.t to other resources

New Annotation by Inference
A method for inference – testing on a new genome

New Function to Disserted Proteins
High level functionality – story of the toxin like proteins
Why domain families? what is wrong with protein classification

Nothing is wrong, But:

- Reducing **false** transitivity.
- Exposing **Mix and Match** evolution
- **Immediate relevance** to **structural** domain-families
- Suggesting evolutionary **‘robust units’**

Why automatic?

Overcoming large **amounts of data**

**Unbiased identification** of new families (even without an identified seed)
EVEREST: A domain families resource
A comparative quality tool for other resources

Automatic / de-novo identification and classification of protein domains in all known sequences

Rigorous evaluation against manually / automated & structurally based domain-family resources

- Scoring methods for a ‘quality control’
- Exposing any (interesting) relationships within ‘the world’ of domains

- Web interactive tool
  www.everest.cs.huji.ac.il
The Modular Nature of Proteins

Method

K6A1 MOUSE

CSKP HUMAN

DLG3 MOUSE

MPP3 HUMAN

- Serine/Threonine protein kinase family active site
- Protein kinase C-terminal domain
- PDZ domain
- SH3 domain
- Guanylate kinase
If we cluster these proteins, assuming transitivity of local alignment scores, we will cluster K6A1_MOUSE with MPP3_HUMAN.
Each BLAST alignment defines two segments.
Clustering Segments

Two similarity measures between segments:

- **Sequence similarity** if they were found together by BLAST
- **Physical overlap** if they are on the same protein, and they intersect
All segments on CSKP_HUMAN defined by alignments with e-score 1e-40 or better:

We collect all Blast value that are < 100! ~14 million values
EVEREST: Process Scheme

**Evolutionary Ensembles of Recurrent Segments**

Pre-process

Iterations

Post-process

Evaluation and tests

Careful transitivity

Putative domains

Putative families

Statistical model

Machine learning

Clustering

Majority voting

Method
Cluster the segments into conservative groups by overlap similarity.

Each group is a **putative domain**

- We apply average linkage hierarchical clustering on the putative domains.
- Creates a binary tree of clusters.
- Each cluster is a **putative domain family**

- Machine learning & Scoring w.r.t. PfamA.
- Choosing good families (intrinsic properties) – training/ disjoin to test.
- Each family modelled by HMM, redefine **EV families**.
- Iteration (3 times from 100K to 25K).
- Jointing HMMs and voting for EV consensus family.
Quality & Evaluation

Comparing with **Pfam**
Pfam is a domain signature DB, manual curation, covers 62% aa, 7500 signatures

**Accuracy** – how well a typical EVEREST domain family scores w.r.t Pfam

Size of the intersection over the size of the union
Scores range from 0 to 1.0 (Jaccard Score)

EV of 10 instances matches Pfam with 10 with only 9 are overlapping

Score: 0.81
Getting Better (accuracy measure)

All Clusters

Score wrt Pfam

~ 2 million

Chosen Clusters

Score wrt Pfam

~ 100,000

Iteration 1 HMMs

Score wrt Pfam

~ 100,000

Iteration 3 HMMs

Score wrt Pfam

~ 25,000

Final EVEREST Families

Score wrt Pfam

~ 13,570
EVEREST - Evaluation vs Reference

- EVEREST is evaluated against reference sets of known families (Pfam, SCOP, CATH)

- Score of EVERSET family w.r.t. Intersecting reference family:
  - size of intersection / size of union

- **Accuracy**
  - Each EVEREST family scored vs. best matching reference
  - Look at score profile across EVEREST families
  - Ignore EVEREST families unknown to reference set

- **Coverage**
  - Each reference family scored vs. best matching EVEREST
  - Look at score profile across interesting subsets of reference set
  - Non-Trivial: family size $\geq 5$
  - Hetero: non-trivial + appearing in hetero-multi-domain proteins
Evaluation –wrt Pfam  EVEREST & ADDA (Holm)

EVEREST - Accuracy

ADDA - Accuracy

EVEREST - Coverage

ADDA - Coverage

13,570

1,800

# families
EVEREST & ADDA
Evaluation vs Pfam

Accuracy vs Pfam

Coverage of Hetero Pfam Families

Hetero >5
Evaluation – Compare w.r.t SCOP manual classification of structural domains

EVEREST - Accuracy

EVEREST - Coverage
EVEREST - Evaluation vs SCOP (family) coverage

Coverage of all SCOP families

Coverage of Non-Trivial SCOP families
Evaluation – Compare wrt CATH /SCOP superfamilies (coverage)
13,569 EV families were defined. Providing Joint HMMs.

Jointly cover 83% of the aa in the SWP DB.

The average (median) size of an EVEREST domain family is 81 (41).

The average (median) length of the domains is 117 (76) aa.

Move to some examples (web based querying)
**Examples: New Functional Annotation**

- **EVEREST family 1017**
- **PF04673 (Polyketide synthesis cyclase)**
- **PF04486 (SchA/CurD like protein)**

- PF04486 has no known function.
- Two of its members are known to be in gene clusters involved in the synthesis of polyketide-based spore pigments.
- Could these two families be considered one?
New Family (1)

- EV02275 is unknown to Pfam
- 54 out of its 55 domains appear 90 positions N-terminal to PF03171 (2OG-Fe(II) oxygenase superfamily)
- Perhaps this is a new domain family?

- PDB 1UOG
  - RED – EVEREST 2275
  - BLUE - PF03171
New domain family (2)

48 proteins – Pesticidial crystal protein cry5Aa
(Insecticidal delta-endotoxin CryVA(a) (Crystaline entomocidal protoxin)

EV covers the 48 proteins of PFAM (and SCOP / CATH) - perfectly

but another EV specifies the family – no OVERLAP and NO structure for this region (609-911)
Two that became one

Examples in Pfam CLANs

PFAM (OLD) Taurine catabolism dioxygenase TauD, TfdA family
Pfam (NEW) a composed entry: **TauD**
Superfamily

- EVEREST family **EV04463** fully covers both PF00465 (Iron-containing alcohol dehydrogenase) and PF01761 (3-dehydroquinate synthase).
- **ENZYME**: PF00465 is EC1.1-
- **ENZYME**: PF01761 is sometimes EC4.6 and sometimes EC1.1
- **SCOP /CATH**: Same superfamily/ Homology group

PDB 1JQA (PF00465)  PDB 1DQS (PF01761)
Alternative Family Definition

Elongation Factor

3 ‘domain family’: All support same proteins

Half C-terminal
SCOP - two adjacent domains (yellow, blue)
CATH – two separated (blue, red) spacer (green)
EVEREST – one domain (pink)
### Protein P-64671

<table>
<thead>
<tr>
<th><strong>Everest ID</strong></th>
<th>P-64671</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>System</strong></td>
<td>SwissProt 40.26</td>
</tr>
<tr>
<td><strong>ID in Source</strong></td>
<td>SC17_YEAST</td>
</tr>
<tr>
<td><strong>Accession number</strong></td>
<td>P52602</td>
</tr>
<tr>
<td><strong>EMBL Protein-ID</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Protein name</strong></td>
<td>Vesicular-fus</td>
</tr>
<tr>
<td><strong>Length in amino acids</strong></td>
<td>291</td>
</tr>
<tr>
<td><strong>Theoretical pI</strong></td>
<td>4.96</td>
</tr>
<tr>
<td><strong>Molecular weight</strong></td>
<td>32757 Da</td>
</tr>
</tbody>
</table>

#### Sequence of protein

<table>
<thead>
<tr>
<th>Position</th>
<th>Amino Acid</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>M</td>
</tr>
<tr>
<td>51</td>
<td>H</td>
</tr>
<tr>
<td>121</td>
<td>E</td>
</tr>
<tr>
<td>131</td>
<td>K</td>
</tr>
<tr>
<td>241</td>
<td>E</td>
</tr>
</tbody>
</table>

#### Keywords

**Swissprot**

- Endoplasmic reticulum, Golgi stack, Protein transport, Transport

**InterPro accession number**

- IP00074

**GO cellular component**

- Golgi apparatus, Cell, Cellular_component, Cytoplasm, Endoplasmic reticulum, Intracellular

**GO molecular function**

- Intracellular transporter, Molecular_function, Protein transporter, Transporter

**GO biological process**

- Biological_process, Cell growth and/or maintenance, Intracellular protein transport, Protein transport, Transport

#### NCBI Taxonomy

- **SUPERKINGDOM** - eukaryota
  - **KINGDOM** - fungi
    - **PHYLUM** - ascomycota
      - **SUBPHYLUM** - saccharomycotina
        - **CLASS** - saccharomycetes
          - **ORDER** - saccharomycetales
            - **FAMILY** - saccharomycetaeae
              - **GENUS** - saccharomyces
                - **SPECIES** - saccharomyces cerevisiae
Evaluate any reference domain resources

Display settings:
Choose sequence databases
Choose domain family systems

Family page header:
General statistics
Download of HMMs
Links to list of domains and to evaluation pages

View EVEREST Family EV00014

Number of proteins in PDB: 4 (representative: 2).
Number of proteins in Swiss-Prot: 16 (representative: 20 (representative: 15).

Download HMMs:
Global-Local HMM:
Global-Global HMM:

- List of domains of EV00014 in tabular form.
- Scoring of EV00014 by families from other systems.
- Scoring of families from other systems by EV00014.
<table>
<thead>
<tr>
<th>Family ID</th>
<th>Count</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>10369</td>
<td>9</td>
<td>25</td>
</tr>
<tr>
<td>12204</td>
<td>9</td>
<td>25</td>
</tr>
<tr>
<td>1875</td>
<td>9</td>
<td>25</td>
</tr>
<tr>
<td>10564</td>
<td>9</td>
<td>25</td>
</tr>
<tr>
<td>2328</td>
<td>5</td>
<td>17</td>
</tr>
<tr>
<td>6667</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>2118</td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>11462</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>11047</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>111787</td>
<td>1</td>
<td>3</td>
</tr>
</tbody>
</table>

Legend:
- Families
- Appearances

Is there any added value for The overlapping EV families?

**EV10564** / 100% - perfect match but 220 aa not 640 aa

**EV01875** / 87% cover / 3 new
### Family color code legend:
Current family always in red
Relationship of current family to other families

**Type refers to relationship between boundaries:**
- **same** = similar boundaries
- **subdomain**
- **superdomain**
- **C-terminal neighbor**
- **N-terminal neighbor**

**Forward** = “how many of the member of the current family participate in the relationship”

**Backward** = “how many of the member of the other family participate in the relationship”

### SCOP Family Neighbors

<table>
<thead>
<tr>
<th>Neighbor</th>
<th>Type</th>
<th>Level</th>
<th>Forward</th>
<th>Backward</th>
</tr>
</thead>
<tbody>
<tr>
<td>d.224.1.2</td>
<td>Same</td>
<td>(2/2)</td>
<td>(2/2)</td>
<td>(2/0)</td>
</tr>
</tbody>
</table>

### PFAM Family Neighbors

<table>
<thead>
<tr>
<th>Neighbor</th>
<th>Type</th>
<th>Level</th>
<th>Forward</th>
<th>Backward</th>
</tr>
</thead>
<tbody>
<tr>
<td>PF01592</td>
<td>Same</td>
<td>(12/13)</td>
<td>(12/14)</td>
<td>(10/1)</td>
</tr>
<tr>
<td>PF01106</td>
<td>C-Term</td>
<td>(5/13)</td>
<td>(5/15)</td>
<td>(4/0)</td>
</tr>
<tr>
<td>PF04324</td>
<td>C-Term</td>
<td>(3/13)</td>
<td>(3/20)</td>
<td>(3/1)</td>
</tr>
</tbody>
</table>

### Download HMMs:
- Global-Local HMM: 
- Global-Global HMM:

- List of domains of EV00014 in tabular form.
- Scoring of EV00014 by families from other systems.
Next Phase:

- Improving EVEREST web
- **Evaluation** of ALL used resources
- Phylogenetic View
- Enrich queries (according to reference Resource)
- Names for EVxxxx
- **Paste** your protein
- Domain **boundaries**

79 proteins
30S ribosomal protein S4
Summary:

- We provide an automated framework for identification and classification of new protein domains
- recovering 60% of difficult known Pfam families.
- Suggests new families for 8% (with > 51% fidelity)
- For 20% we suggest a new view on domain families

- Manual inspection of families scoring low w.r.t. Pfam suggested that many of those are valid families.

- Enabling inspection of EVEREST families and additional resources in http://www.everest.cs.huji.ac.il
Annotation & Inference
New genomes, New functions

EVEREST

Automatic
(no pre-knowledge)

Partition to ‘domains’
(no transitivity)

Robustness
(evaluate w.r.t others)

Having Function
Experiments
Literature
Expert view

No Function
New genomes
No similarity
No evidence
Domain families by EVEREST
Automatic identification of Protein Domain
Performance and analysis w.r.t to other resources

**New Annotation by Inference**
A method for inference – testing on a new genome-the BEE

**New Function to Disserted Proteins**
High level functionality – story of the toxin like proteins
### Honey Bee
**The brain & complex neuronal behavior**

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>C Elegans</td>
<td>19,000</td>
</tr>
<tr>
<td>(worm)</td>
<td></td>
</tr>
<tr>
<td>Miniat. Wasp</td>
<td>10,000</td>
</tr>
<tr>
<td>Drosophila</td>
<td>14,000</td>
</tr>
<tr>
<td>(fruit fly)</td>
<td></td>
</tr>
<tr>
<td><strong>Apis</strong></td>
<td><strong>10,000</strong></td>
</tr>
<tr>
<td>(honey bee)</td>
<td></td>
</tr>
<tr>
<td><strong>Homo Sapiens</strong></td>
<td><strong>25,000</strong></td>
</tr>
</tbody>
</table>

The number of neurons or genes is not indicative for the brain and behavior complexity.

**The makeup of a social behaving insect**
Honey bee genome recently sequenced: ~200 MB (by HGSC at Baylor College of Medicine)

10,157 predicted ORFs

- Produce a hierarchical (functional) organization of the bee proteome
- Annotate the bee sequences
- Systematically find putative instances of
  - Bee gene-loss events
  - Bee-specific paralogs
  - Bee-specific functionality
  - Mis-predicted genes (FN/FP)
ProtoNet classifications
The Principles: A reminder

• Unsupervised
• Only sequence information as input
• All proteins involved (incl. hypothetical...)
• Family definition is hierarchical
• Only based on statistical significance of the similarity score

• Clustering process after ALL mutual ‘distance’ information is computed
  (Blast of All against All for 120 K proteins, E=100)

Evaluation vs InterPro, GO etc
Pfam, Prosite, SMART, PRINTS, SCOP, CATH...

www.protonet.cs.huji.ac.il
Clustering Method

First, each protein is considered a singleton (a cluster of its own).
● Next, we iteratively merge the pairs of clusters
● We choose to merge the ‘most similar’ pair of clusters.
Clustering Method
The clustering process gradually generates a tree of clusters

Merging Scores

Pruning:
Compact the tree to 12% of its size without
Reduction in performance (w.r.t. InterPro)
quality..
ProtoNet Hierarchical organization

Protein database:
SwissProt \(~133,000\) proteins –
Testing the ‘Matching Score’ for InterPro (combining all high – quality domain based / structure base / knowledge based)

\[
\text{score}(C, S) = \frac{|C \cap S|}{|C \cup S|}
\]
Annotation Inference for proteins in clusters

C- cluster C; K - keyword

**Annotation Score** \( AS(C,K) = \) specificity\(^2 \times\) sensitivity = 0.25

\[
\left( \frac{TP}{TP + FP} \right)^2 \times \frac{TP}{TP + FN}
\]

\( TP \) is the proteins in C that have the keyword K
\( FN \) is the proteins not in C that have the keyword K
\( FP \) is proteins in C that do not have the keyword K.

The high-confidence annotation threshold

ProtoNet > 20 pr
Method for the Bee
Hierarchical organization

Protein database (200,000 pr) •
Predicted bee protein set: 10,157 pr –
SwissProt (without bee) – ~133,000 proteins. –
Drosophila proteome (insect) – 20,730 pr. –
mouse proteome (UniProt) – 35,199 pr. –

All vs all BLAST •
Clustering •
Tree chopping •
Tree pruning •

similarity

low

high
ProtoBee: results

Clustered into 5095 families (out of 18,500)

Mouse
185 (151)

Fly
927 (733)

Other
6779 (2539)

Unique
159 (143)

Other
365 (302)

Fly
707 (500)

Mouse
439 (389)

Other
596 (338)

Unique
185 (151)
For each cluster, **calculate its annotations**. Each annotation is required to:

(a) be assigned > **75%** of the proteins in the cluster  
(b) achieve **p-value <= 0.001** (hypergeometric distribution).

Only clusters with > **5 proteins** are considered  
For each bee protein, assign to it the annotations of its cluster and all parents.
Annotation summary

- Combined
- GO
- GO MF
- GO BP
- InterPro
- SwissProt
- EC

Bar chart showing:
- Hydrolase activity: 8000 proteins
- Nucleotide metabolism: 7000 proteins
- Transport: 6000 proteins
- Protein metabolism: 5000 proteins
- Transferase activity: 4000 proteins
- Cell communication: 3000 proteins
- Signal transduction: 2000 proteins
- Oxidoreductase activity: 1000 proteins
- Transcription: 500 proteins
- Receptor activity: 100 proteins

Y-axis: Number of proteins
X-axis: Number of proteins
How good is this method?

**Pros** (assuming negligible transitivity):

Any kind of external information source can be used for annotation.

“Robustness” reduces chance of false positives. –

Potentially links biological properties to localized sequence features.

**Cons:**

Incorrect transitivity due to multiple domains. –

Not as sensitive/specific as motif-based methods. –
Results overview

Clusters organized into 18,936 trees (roots).
5095 roots contain bee proteins.

Annotation: 70% of proteins are annotated (InterProScan covers ~72-78%).

Interesting biological information on the evolution of the bee relative to other insects (different talk)
Annotation & Inference
New genomes, New functions

ProtoBee

Having Function
- Annotation Score (high confidence)
- Clusters leading to Retesting ORFs

No Function
- New genomes
  - No similarity
  - No evidence

Having Function
- Experiments
- Literature
- Expert view

May 2006