

The Protein Space

From Sequence to Structure to Function

Michal Linial

Institute of Life Sciences
The Hebrew University
Jerusalem, Israel





Protein Sequences

1,000,000 pr
(static)

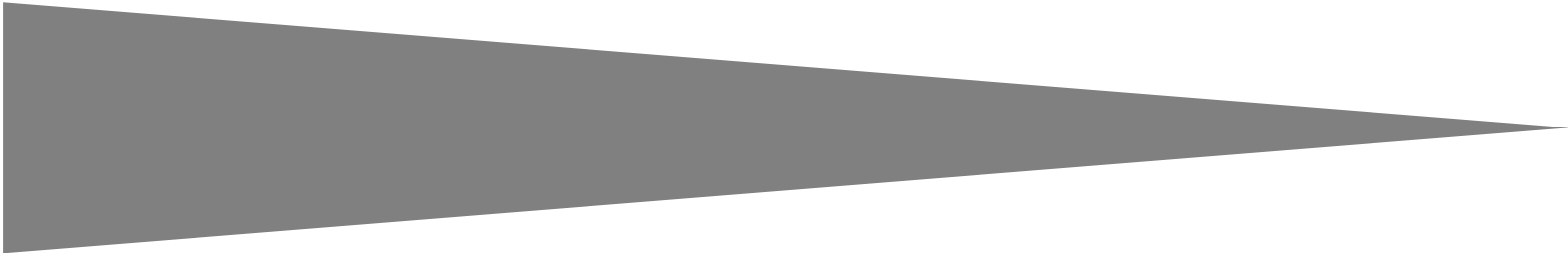
Protein Variants

10,000,000 pr
(dynamic)

Exon combinations, post-translation modification, p-p interaction...

Protein Function

?????



A link between sequence, structure and function

Protein **structures** are much more conserved than protein **sequences**

Proteins of identical (similar) **Structure** tend to have identical (similar) **function**

Try to assign function based on structural characterization

Extract structural information from sequence alone
(The Holy Grail)

The Goals

Collaboration with CESG

- THE GOALS:**
1. Reducing the proteins space to ~10,000-15,000 clusters (singletons ??).
 2. Constructing a 'best map' for *A. Thaliana*
 3. Functional annotations for AraNet (protein net based on *A.thaliana* proteins)
 4. Ranking *SG* targets for *Arabidopsis Thaliana*
 5. Cope with accumulation of *SG* structures

The Scheme of the talk

THE TASK: Construct a map of the protein space

ProtoClass -rationale and concept

ProtoNet in brief

AND BEYOND: Functional roadmap in ProtoClass

Biological examples

THE PURPOSE: New superfamilies for Structural Genomics

ProTarget - ranked list of proteins

THE APPLICATION: **AraNet** - in view of SG

QUO VADIS: Quality assessment of clusters for FG

vis-à-vis InterPro, SCOP, FSSP etc

ProtoClass - Set of automatic classifications of all proteins

Seeking statistically significant regularities (clusters)
Reconstruct the 'geometry' of the sequence space



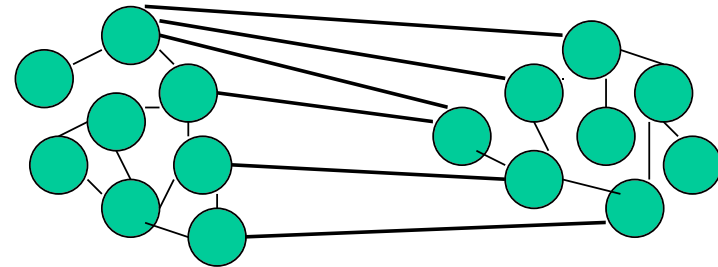
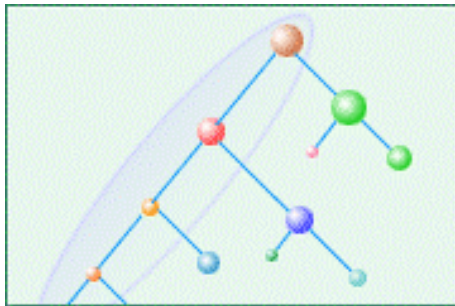
Guiding principle

Homologous proteins evolved from common ancestor protein

Homology is a transitive relation that can be deduced based on statistical similarities

Large-Scale Clustering - some theoretical background

- The proteins form a large metric space. This leads to the theory of metric embeddings. (Geometric approach).



- A combinatorial perspective: Proteins are the vertices of a large graph. Seek proper rules for merging related clusters.

ProtoClass - main features

Pairwise distances (**all against all**)

Includes all **SwissProt** proteins

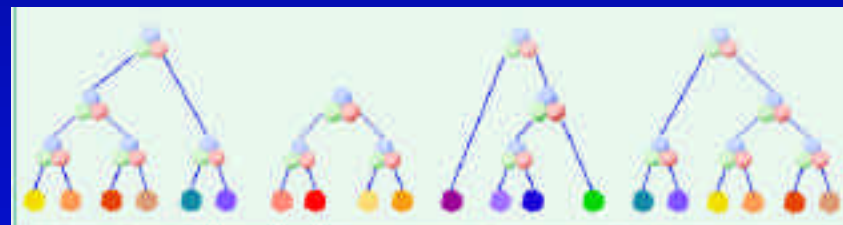
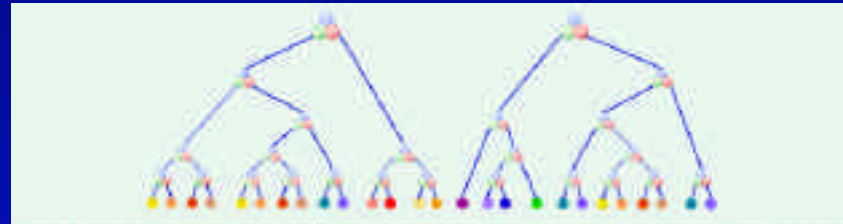
Graph based

Unsupervised and **automated**

Hierarchical

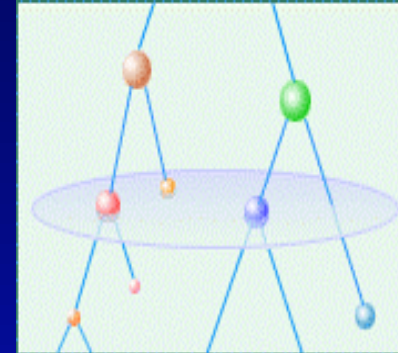
A clustering **algorithm** based on a 'merging score'

Bottom-up clustering



ProtoClass - Set of classifications of all proteins

ProtoClass systems generate graphs and maps that yield views at any levels of granularity.



ProtoMap



release May 1997

ProtoNet - A (arithmetic)

release July 2002

ProtoNet - G (geometric)

release July 2002

ProtoNet - H (harmonic)

release July 2002

ProtoNet - G50

release Dec 2002

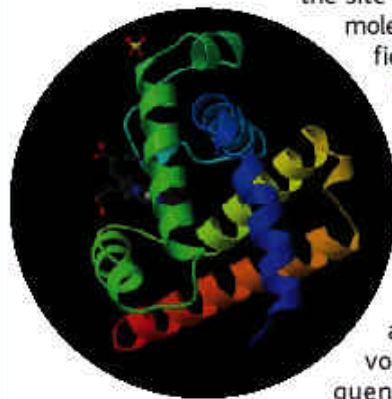
ProtoNet - A50

release Dec 2002

DATABASE

Parse Protein Pedigrees

A new Web site from Hebrew University in Jerusalem aims to simplify the analysis of protein structure and function. Along with the usual sequence information, ProtoNet automatically clusters proteins by similarity, creating a family tree that allows researchers to compare individual proteins or related groups. For more than 100,000 proteins,



the site holds a data card that lists each molecule's amino acid sequence, identifies functional regions, and charts the taxonomy of the organism it comes from. You can compare each protein to other members of its immediate family or climb up the tree to contrast different groups, which might help deduce the function of mystery molecules or tease out evolutionary trends. If you don't find your favorite protein here, submit its sequence to find out how it fits into known clusters.

www.protonet.cs.huji.ac.il/protonet/index.php

Science 298 : p329 (11 October 2002)

ProtoNet - in brief

ProtoNet.cs.huji.ac.il
automatic hierarchical classification of proteins

Version 1.4

main page
search
classify your protein
introduction
methods
guided tour

related links
ProtoNet team
help
feedback

search _____

Individual protein

Get protein card
Returns information about a protein.

Protein Sequence Alignment
Displays sequence alignment for two proteins.

Check protein in cluster
Testing whether a protein belongs to a given cluster.

Vertical perspective

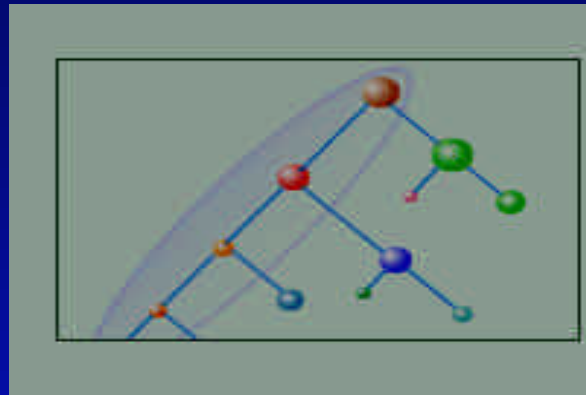
Get cluster card
Returns information about a ProtoNet cluster.

www.protonet.cs.huji.ac.il

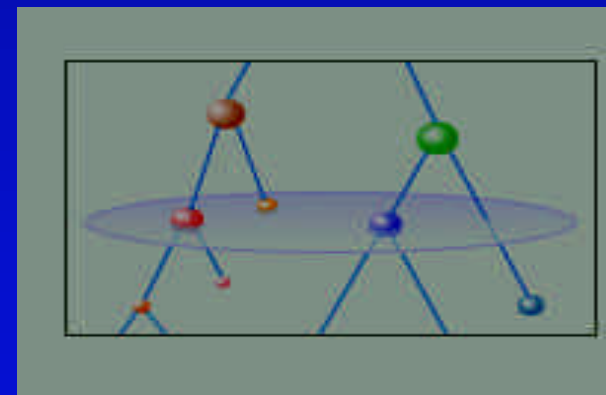
ProtoNet Perspectives



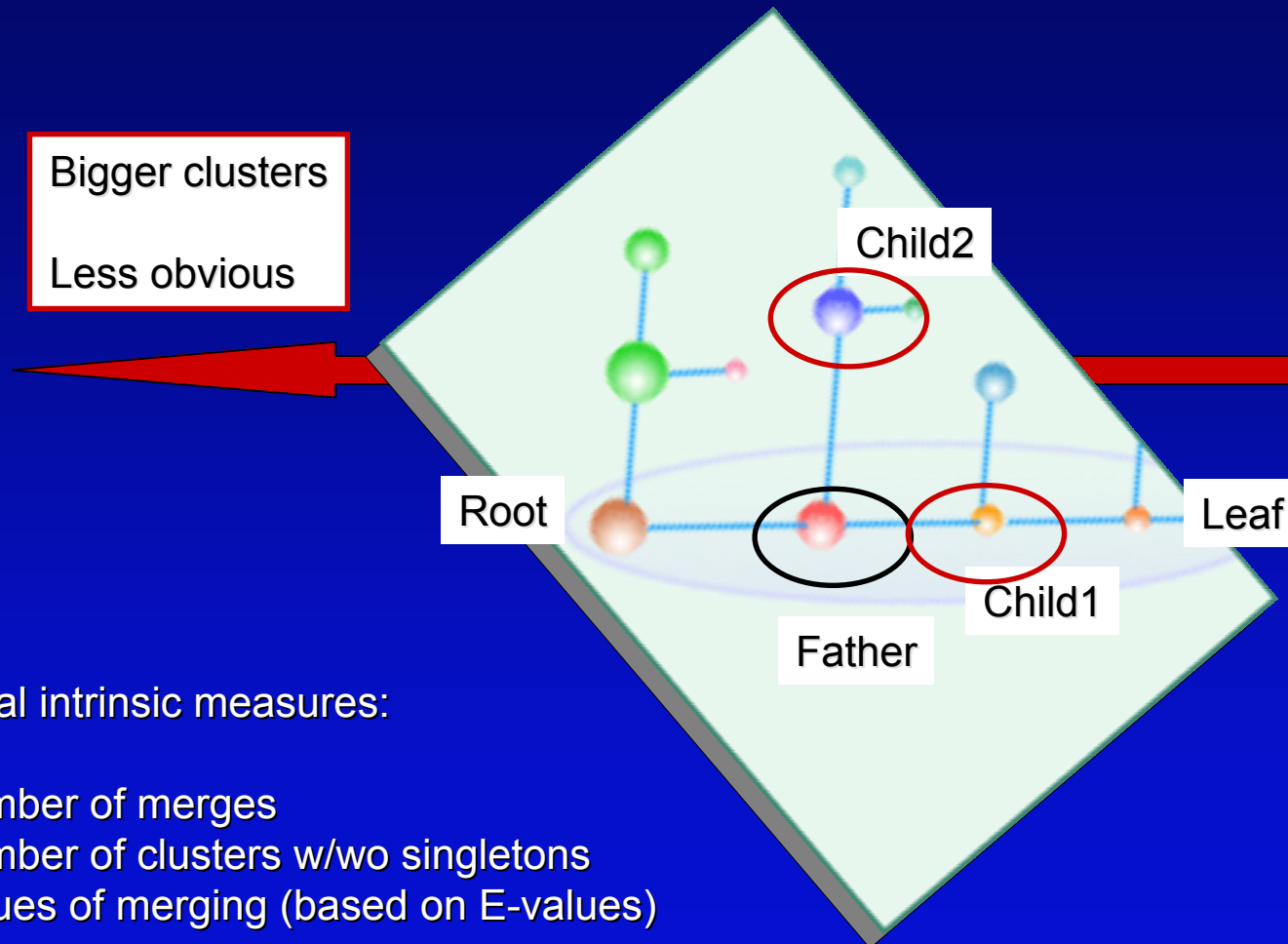
Clustering Chain (to the root - subfamilies)



Horizontal (to the correct level - maps)



Clustering Chain



Several intrinsic measures:

- Number of merges
- Number of clusters w/wo singletons
- Values of merging (based on E-values)