Consistent Triplets in Graph Clustering for Protein Sequence Analysis

HwaSeob Joseph Yun

Department of Computer Science
Rutgers University

April 26, 2006

Outline

1. Motivating biological problem
2. Prior work on graph clustering
3. **Consistent Triplets** for clustering
4. Results: comparisons to authoritative curated clusters

Clustering Protein Sequences

- **Protein**: a sequence of amino acids, which determines a **unique 3-dimensional structure** capable of various functional roles.
- **Paralog**: closely functionally related proteins **within** a genome resulting from duplication events.
- **Problem:** Experimental test of new clustering to find paralog candidates.

3 / 41

Pairwise and Multiple Sequence Alignment

- For a given cluster, multiple sequence alignment is the **best validation** test.
- Multiple sequence alignment is **very slow**: it cannot be used to find clusters.
- All known bioinformatics methods use **only pairwise** sequence alignment for clustering.

4 / 41

Protein Sequence Similarity Score

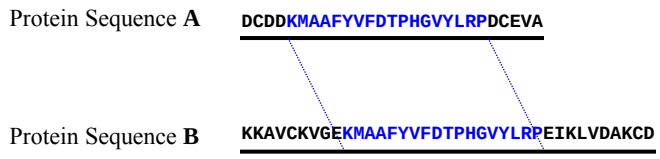
- There are several score functions which are calculated from pairwise sequence alignment: % of identities, % of gaps, E-value.

```
MSCFVTEKKAVCKVGEKMAAFYVFDTPHGVYLRPEIKLVDDWIKVAHRGDDK
                    ||||||||||||||||||||+|+|||||||||
                    MAAFYVFDTPHGVYLRPEIKLIDEWIKVAHRGDGGG
```

- My method uses E-value which is an estimation of a probability that the analyzed database matches may have occurred just by chance.

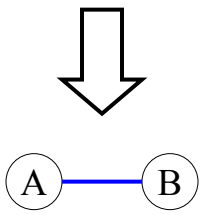
5 / 41

Pairwise Sequence Alignment in Bioinformatics Clustering



- Pairwise sequence alignment is used only to measure similarity for pairs of protein sequences:

  to build a graph in which protein sequences are nodes, and edges for pairs of protein sequences with a high similarity score.



6 / 41

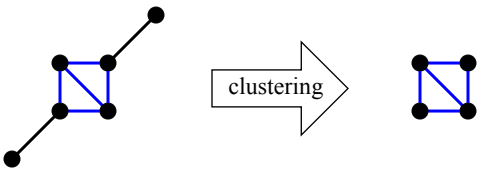
Multiple Sequence Alignment (MSA) Approximation: Triplets in Clusters

- Almost 10 years ago, researchers in bioinformatics found it necessary to find an approximation of MSA, which can be used for protein sequence clustering.
- **COG & KOG** [Koonin, et al. A Genomic Perspective on Protein Families. *Science*, 1997.]

7 / 41

MSA Approximation: the Basic Idea

- Use the standard graph with edges which are related to high score values, and reduce it so that every edge is involved in at least one connected **triplet**. And use a standard graph clustering technique over this reduced graph.



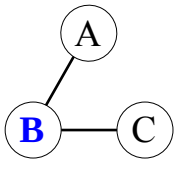
8 / 41

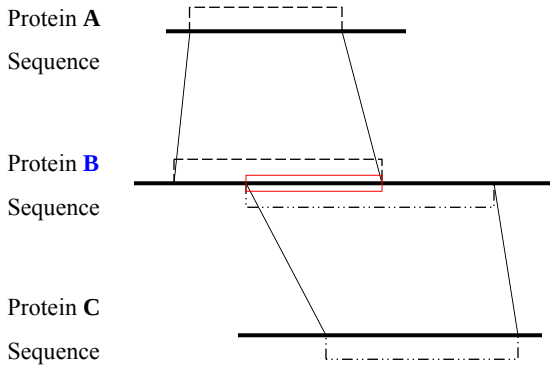
Transitivity Property

- Such reduction was motivated by the idea to keep transitivity property of **connectivity within a cluster**, because it was found experimentally that this property represents better evolutionary similarity.
- [Koonin, et al. The structure of the protein universe and genome evolution. *Nature*, 2002.]

Novelty in my research

Automating extraction of connected triplet supported by **significant overlap**.



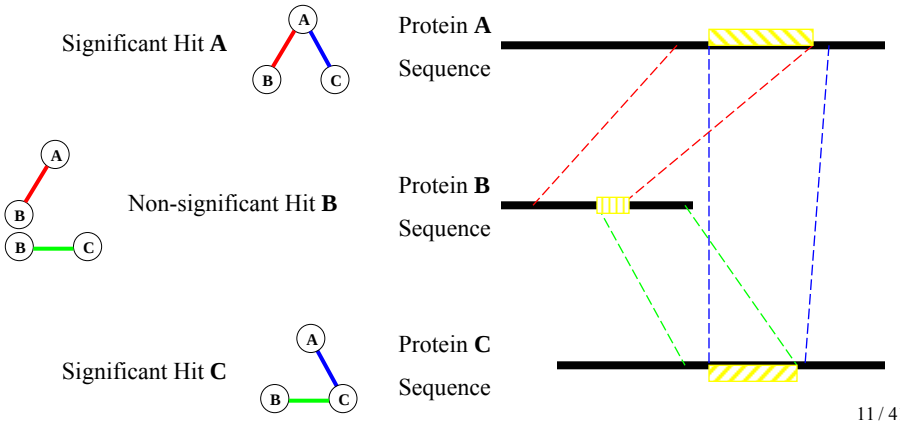


Protein B is called a **significant hit** for A & C

if it has edges (B, A) and (B, C), and $Overlap_B(A, C) \geq L_0$

Necessity to check the Significant Hit property for every node in connected triplet

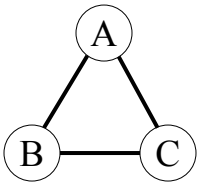
Example of Inconsistent Triplet
due to Lack of a Significant Hit



11 / 41

Consistent Triplet (CT)

- Each node in a connected triplet is a significant hit for the other two.



1. A is a significant hit for B and C.
2. B is a significant hit for C and A.
3. C is a significant hit for A and B.

12 / 41

Criterion to Find CT-cluster: the specific novelty of my research

- ***W*** = the set of protein sequences in a **genome**
- Protein sequence ***i*** $\in$ subset ***H*** $\subseteq$ whole set ***W***
- $\pi(i, H)$ = number of consistent triplets within H for ***i***
- Score function $F(H) = \min_{i \in H} \pi(i, H)$
- **Problem**: Find $\max_{H \in 2^W - \varnothing} F(H)$
- The solution cluster H^* guarantees that every protein in H^* is involved in at least $F(H^*)$ number of CTs.

13 / 41

The Basic Algorithm to Find CT-cluster

The algorithm is the following iterative procedure:

1. Find $F(H^*)$ in G_i (i = 1, original graph) and build the subgraph G_{i+1} on $G_i - H^*$ nodes. Keep H^* as a CT-cluster.

2. If $|G_i - H^*|$ does not include any CT, stop; otherwise repeat 1.

14 / 41

How to find the global maximum $F(H^*)$?

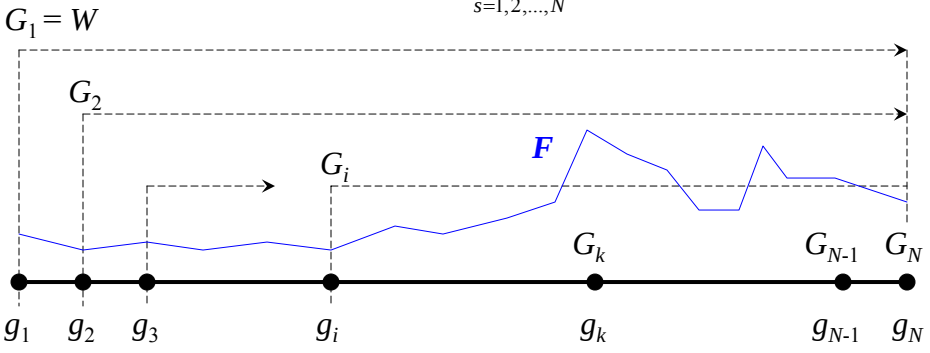
- Mullat's shelling procedure is used: two sequences of sets and their score function values G & F are built.
  - Step 1. $g_1 \in G_1 = W$, $F(G_1) = \min_{s \in W} \pi(s, W) = \pi(g_1, W) = F_1$.
  - Step 2. $g_2 \in G_2 = G_1 - g_1$, $F(G_2) = \min_{s \in G_2} \pi(s, G_2) = \pi(g_2, G_2) = F_2$.
  - Step i. $g_i \in G_i = G_{i-1} - g_{i-1}$, $F(G_i) = \min_{s \in G_i} \pi(s, G_i) = \pi(g_i, G_i) = F_i$.
  - Step N. $g_N \in G_N = \{g_N\} = G_{N-1} - g_{N-1}$, $F(G_N) = \pi(g_N, G_N) = F_N$.

[Mullat, Extremal Subsystem Of Monotone Systems. *Automation and Remote Control*, 1976]

15 / 41

Shelling Procedure: F and G

- Find the smallest index k
  in the sequence $F = \langle F_1, F_2, \ldots, F_N \rangle$
  which satisfies $F_k = \max_{s=1,2,\ldots,N} F_s$.



16 / 41

Recursive Decomposition to speed up the basic algorithm to find CT-clusters

- **CT-subgraph**: any subgraph where each edge is involved in at least one CT.
- The basic algorithm works only on CT-subgraphs.
- Complexity for finding $F(H^*)$ is $O(n^4)$ where n is a number of nodes in the considered CT-subgraph.

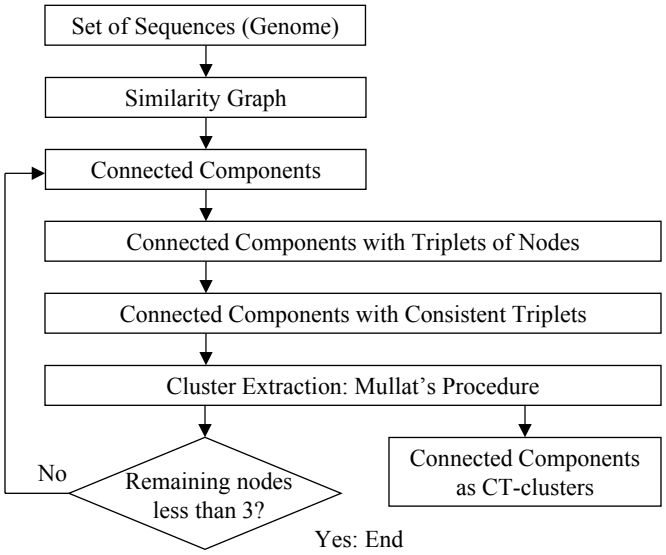
17 / 41

Recursive Decomposition

- Initially, the procedure finds all maximal connected CT-subgraphs from the original graph.
- After finding $F(H^*)$ in all CT-subgraphs, and keeping them as CT-clusters, procedure searches the rest of the CT-subgraphs to find new (smaller) CT-subgraphs.

18 / 41

CT-clustering Process Overview



19 / 41

Evaluation of CT-clustering

- Comparison with COG/KOG and KEGG **http://www.cs.rutgers.edu/~seabee/**
- Sensitivity analysis of clusters for range of thresholds (similarity & overlap)
- Two cases of specific biological function subclasses

20 / 41

Paralog Candidates by CT-clustering

Unicellular Clusters (COG)		
Group	species	
Archaea	13	Afu Hbs Mac Mth Mja Mka Tac Tvo Pho Pab Pya Sso Ape
Eukaryota	3	Sce Spo Ecu
Bacteria	10	Aae Tma Ctr Cpn Tpa Bbu Syn Nos Fnu Dra
Actino bacteria	4	Cgl Mtu MtC Mle
Gramplus	12	Cac Lla Spy Spn Sau Lin Bsu Bha Uur Mpu Mpn Mge
gamma	11	Eco EcZ Ecs Ype Sty Buc Vch Pae Hin Pmu Xfa
Proteo bacteria	6	Nme NmA Rso Hpy jHp Cje
alpha	7	Atu Sme Bme Mlo Ccr Rpr Rco
Total	66	

Eukaryotic Clusters (KOG)		
Code	*Name*	Abbreviation
A	*Arabidopsis thaliana* (thale cress)	***ath***
C	*Caenorhabditis elegans* (worm)	***cel***
D	*Drosophila melanogaster* (fruit fly)	***dme***
H	*Homo sapiens* (human)	***hsa***
Y	*Saccharomyces cerevisiae* (baker yeast)	***sce***
P	*Schizosaccharomyces pombe* (fission yeast)	***spo***
E	*Encephalitozoon cuniculi* (Microsporidia)	***ecu***

Clustered by

HwaSeob Joseph Yun
seabee@cs.rutgers.edu

- Species: **Homo sapiens**
  Homo sapiens has total 38,638 sequences from KOG-FTP site.
  Total number of sequences for all 7 KOG organisms = 112,920

- Parameters used for this clustering:
  e_1 = e-40: E-value threshold for BLAST
  a_1 = 20 residues: minimum overlap between 2 alignments
  Score = guaranteed # of consistent triplets per node

- Shelling 1: 366 (sequences) all from 1 KOG, **Score** = **49,768**
  Node Degree: 323 = min, 365 = max, 363.79 = avg, 1 CT-cluster

  Shelling 2: 117 (sequences) from 2 KOGs, **Score** = **6,670**
  Node Degree: 116 = min, 116 = max, 116.00 = avg, 1 CT-cluster

  Shelling 3: 82 (sequences) from 10 KOGs, **Score** = **1,674**
  Node Degree: 63 = min, 81 = max, 77.00 = avg, 1 CT-cluster

  Shelling 4: 39 (sequences) all from 1 KOG, **Score** = **703**
  Node Degree: 38 = min, 38 = max, 38.00 = avg, 1 CT-cluster

  .....

Cluster 1 of Homo Sapiens

1. Hs13375999 [R] **KOG1721** (498) FOG: Zn-finger
2. Hs7657705 [R] **KOG1721** (417) FOG: Zn-finger
3. Hs21536374 [R] **KOG1721** (310) FOG: Zn-finger
4. Hs20304091 [R] **KOG1721** (292) FOG: Zn-finger
5. Hs15147236 [R] **KOG1721** (316) FOG: Zn-finger
6. Hs21687161 [R] **KOG1721** (306) FOG: Zn-finger
7. Hs22043109 [R] **KOG1721** (914) FOG: Zn-finger
8. Hs22056383 [R] **KOG1721** (349) FOG: Zn-finger
9. Hs22054077 [R] **KOG1721** (1445) FOG: Zn-finger
10. Hs14731015 [R] **KOG1721** (725) FOG: Zn-finger
11. Hs22054039 [R] **KOG1721** (306) FOG: Zn-finger
12. Hs20542862 [R] **KOG1721** (642) FOG: Zn-finger

.....

361. Hs22057914 [R] **KOG1721** (464) FOG: Zn-finger
362. Hs22051365_1 [R] **KOG1721** (824) FOG: Zn-finger
363. Hs17482702_2 [R] **KOG1721** (721) FOG: Zn-finger
364. Hs20471405 [R] **KOG1721** (456) FOG: Zn-finger
365. Hs20471407 [R] **KOG1721** (519) FOG: Zn-finger
366. Hs21314662 [R] **KOG1721** (573) FOG: Zn-finger

CT-clusters from Homo sapiens

.....

Shelling 9: 54 (sequences) from 6 KOGs, **Score** = **300**
Node Degree: 25 = min, 27 = max, 25.89 = avg, 2 CT-clusters

Shelling 10: 30 (sequences) from 8 KOGs, **Score** = **290**
Node Degree: 25 = min, 29 = max, 28.60 = avg, 1 CT-cluster

Shelling 11: 74 (sequences) from 5 KOGs, **Score** = **253**
Node Degree: 23 = min, 25 = max, 23.59 = avg, 3 CT-clusters

Shelling 12: 69 (sequences) from 4 KOGs, **Score** = **231**
Node Degree: 22 = min, 22 = max, 22.00 = avg, 3 CT-clusters

.....

Shelling 9 of Homo sapiens

Total 54 sequences in this shelling.
54 in 6 KOG and 0 not classified by KOG.
Node Degree: 25 = min, 27 = max, 25.89 = avg,
2 CT-clusters. Most common functional categories in order:

28 [T] CELLULAR PROCESSES AND SIGNALING : Signal transduction mechanisms
26 [P] METABOLISM : Inorganic ion transport and metabolism

CT-cluster # 1/2
1. Hs4503859 [T] **KOG3642** (456) GABA receptor
2. Hs12707558 [T] **KOG3642** (393) GABA receptor
3. Hs4557601 [T] **KOG3642** (451) GABA receptor
.....
Hs4557611 [T] **KOG3642** (467) GABA receptor
Hs4503861 [T] **KOG3642** (462) GABA receptor
Hs18558331 [T] **KOG3642** (465) GABA receptor
Hs7657106 [T] **KOG3643** (440) GABA receptor
.....
Hs4503871 [T] **KOG3643** (465) GABA receptor
Hs12548785 [T] **KOG3643** (512) GABA receptor
Hs4504021 [T] **KOG3644** (452) Ligand-gated ion channel
Hs20469202 [T] **KOG3644** (464) Ligand-gated ion channel
.....
28. Hs4504019 [T] **KOG3644** (449) Ligand-gated ion channel

28 nodes with Node Degree: 25 = min, 27 = max, 26.71 = avg

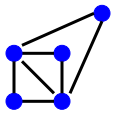
25 / 41

CT-cluster # 2/2
1. Hs13994232 [P] **KOG1545** (456) Voltage-gated shaker-like K+ channel KCNA
2. Hs4504817 [P] **KOG1545** (653) Voltage-gated shaker-like K+ channel KCNA
3. Hs5031819 [P] **KOG1545** (511) Voltage-gated shaker-like K+ channel KCNA
4. Hs4504819 [P] **KOG1545** (585) Voltage-gated shaker-like K+ channel KCNA
5. Hs4504821 [P] **KOG1545** (529) Voltage-gated shaker-like K+ channel KCNA
6. Hs4826782 [P] **KOG1545** (499) Voltage-gated shaker-like K+ channel KCNA
7. Hs4504815 [P] **KOG1545** (523) Voltage-gated shaker-like K+ channel KCNA
8. Hs4557685 [P] **KOG1545** (495) Voltage-gated shaker-like K+ channel KCNA
9. Hs13648551 [P] **KOG1545** (613) Voltage-gated shaker-like K+ channel KCNA
10. Hs19424136 [P] **KOG3713** (545) Voltage-gated K+ channel KCNB/KCNC
11. Hs4758622 [P] **KOG3713** (806) Voltage-gated K+ channel KCNB/KCNC
12. Hs20127427 [P] **KOG3713** (491) Voltage-gated K+ channel KCNB/KCNC
13. Hs21217561 [P] **KOG3713** (613) Voltage-gated K+ channel KCNB/KCNC
14. Hs21217563 [P] **KOG3713** (638) Voltage-gated K+ channel KCNB/KCNC
15. Hs22047567 [P] **KOG3713** (911) Voltage-gated K+ channel KCNB/KCNC
16. Hs4826784 [P] **KOG3713** (858) Voltage-gated K+ channel KCNB/KCNC
17. Hs4826786 [P] **KOG3713** (511) Voltage-gated K+ channel KCNB/KCNC
18. Hs4826790 [P] **KOG3713** (582) Voltage-gated K+ channel KCNB/KCNC
19. Hs19071574 [P] **KOG3713** (436) Voltage-gated K+ channel KCNB/KCNC
20. Hs13492973 [P] **KOG3713** (526) Voltage-gated K+ channel KCNB/KCNC
21. Hs20070166 [P] **KOG3713** (494) Voltage-gated K+ channel KCNB/KCNC
22. Hs14782759 [P] **KOG3713** (477) Voltage-gated K+ channel KCNB/KCNC
23. HsM4826792 [P] **KOG4390** (647) Voltage-gated A-type K+ channel KCND
24. Hs9789987 [P] **KOG4390** (630) Voltage-gated A-type K+ channel KCND
25. Hs4826794 [P] **KOG4390** (655) Voltage-gated A-type K+ channel KCND
26. Hs21361266 [P] **KOG4390** (647) Voltage-gated A-type K+ channel KCND

26 nodes with Node Degree: 25 = min, 25 = max, 25.00 = avg

26 / 41

CT-cluster # 1/825
1. Hs5174755 [R] **KOG3173** (213) Predicted Zn-finger protein
2. HsM9506853 [R] **KOG3173** (186) Predicted Zn-finger protein
3. Hs21359918 [R] **KOG3173** (208) Predicted Zn-finger protein
4. Hs18589968 [R] **KOG3173** (167) Predicted Zn-finger protein
5. Hs14739640 [R] **KOG3173** (159) Predicted Zn-finger protein
5 nodes with Node Degree: 2 = min, 4 = max, 2.40 = avg



CT-cluster # 5/825
1. Hs17466327 [VR] **KOG0184** (124) H Immunoglobulin heavy chain like protein
2. Hs22055964 [VR] **KOG0184** (500) H Immunoglobulin heavy chain like protein
3. Hs22046704 [VR] **KOG0184** (120) H Immunoglobulin heavy chain like protein
3 nodes with Node Degree: 2 = min, 2 = max, 2.00 = avg

CT-cluster # 6/825
1. Hs20468144 [O] **KOG1734** (328) Predicted RING-containing E3 ubiquitin ligase
2. HsM8922863 [O] **KOG1734** (152) Predicted RING-containing E3 ubiquitin ligase
3. Hs21361732 [O] **KOG1734** (327) Predicted RING-containing E3 ubiquitin ligase
3 nodes with Node Degree: 2 = min, 2 = max, 2.00 = avg

KEGG vs. CT-clustering

KEGG	**CT-clustering**
High density set of protein sequences with similarity scores only.	High density set of consistent triplets with similarity and alignment overlap information.
Maximum cliques as clusters	CT-clusters

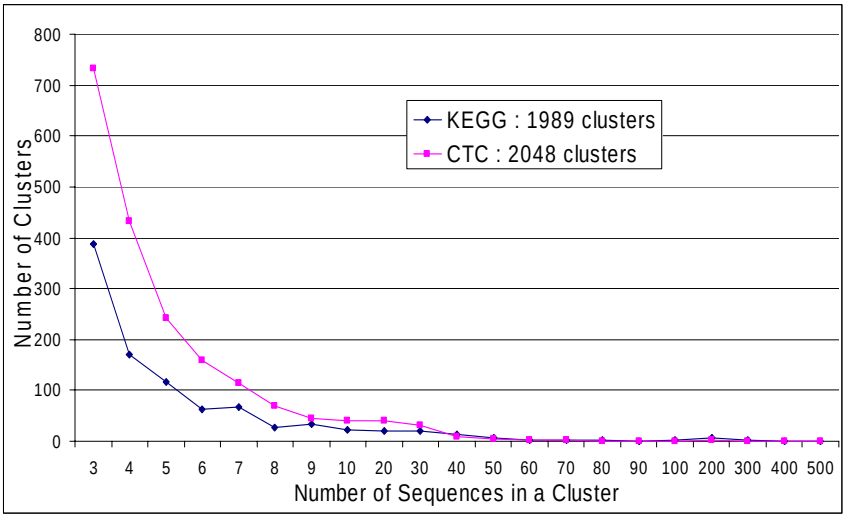
[Kanehisa, et al. Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* , 2000.]

KEGG vs. CT-clustering: Results

Homo sapiens	**KEGG**	**CT-cluster**
Total number of **sequences** in clusters	11,667	12,921
Total number of **clusters**	1,989	2,048
Average size of clusters (standard deviation)	5.8 (17.7)	6.3 (11.4)
Ratio of clusters with 1 or 2 complementary classification	1,791 **90.0%**	1,897 **92.6%**

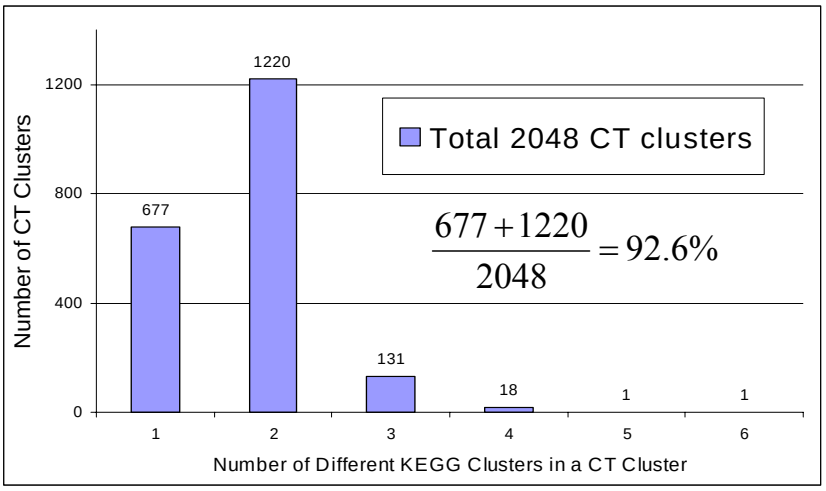
29 / 41

KEGG vs. CT-clustering: Cluster Size Distribution



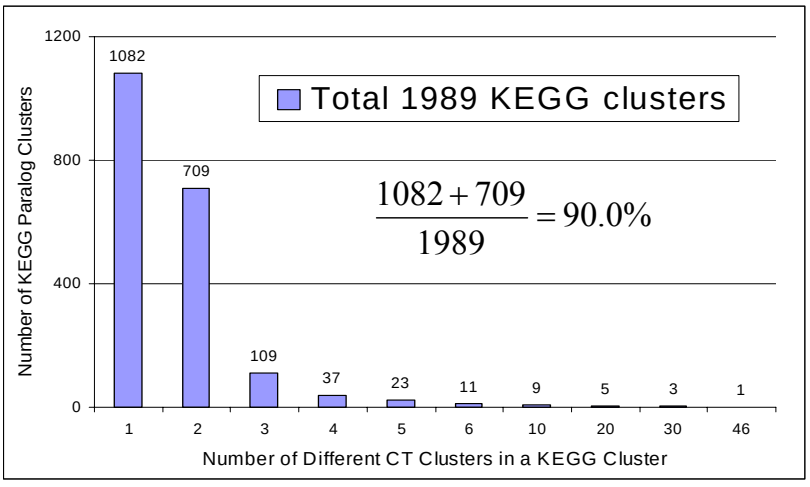
30 / 41

KEGG vs. CT-clustering: Consistency of CT-clusters



31 / 41

KEGG vs. CT-clustering: Consistency of KEGG clusters



32 / 41

Robustness of CT-clustering

Homo sapiens 38,638 proteins		Alignment Overlap Threshold		
		20 amino acids	40 amino acids	60 amino acids
Pairwise Sequence Similarity Score Threshold	10^{-40}	1764 clusters 9842 proteins	1753 clusters 9807 proteins	1745 clusters 9774 proteins
	10^{-60}	1487 clusters 7673 proteins	1481 clusters 7650 proteins	1476 clusters 7626 proteins
	10^{-80}	1251 clusters 6108 proteins	1247 clusters 6094 proteins	1241 clusters 6075 proteins

Analysis of Specific Subclass

- **Transport proteins**
  - A protein that transports a molecule within a cell.
  - Well-known to biologists, naturally forming clusters of paralogous proteins.

Total number of transport proteins in Human genome (KOG)	343 sequences
Identified as 32 CT-clusters with transport proteins only	177 sequences
Identified as 4 CT-clusters with other kinds of proteins	20 sequences
Not included in any CT-clusters	146 sequences

CT-cluster of Transport Proteins

CT-cluster # 9/32
1. Hs20270383 [E] **KOG1287** (470) Amino acid transporters
2. HsM4507055 [E] **KOG1287** (511) Amino acid transporters
3. Hs19923170 [E] **KOG1287** (507) Amino acid transporters
4. Hs9790235 [E] **KOG1287** (523) Amino acid transporters
5. Hs7657683 [E] **KOG1287** (501) Amino acid transporters
6. Hs6912670 [E] **KOG1287** (535) Amino acid transporters
7. Hs4507053 [E] **KOG1287** (515) Amino acid transporters
8. Hs21361563 [E] **KOG1287** (511) Amino acid transporters
9. Hs7657591 [E] **KOG1287** (487) Amino acid transporters

9 nodes with Node Degree: 8 = min, 8 = max, 8.00 = avg

Specific Subclass 2

- **Transcription mechanism**
  - Transcription is a process of copying the DNA-based genetic code to make corresponding messenger RNA.
  - Protein STAT: Signal Transducer and Activator of Transcription in Homo sapiens.
  - Known to be related to dental and heart diseases.
  - KOG has recognized 12 STAT proteins as one cluster.
  - Can CT-clustering identify them as a cluster?

CT-cluster of STAT proteins

1. HsM4507257 [KT] **KOG3667** (794) STAT protein
2. HsM6912688 [KT] **KOG3667** (787) STAT protein
3. Hs21536301 [KT] **KOG3667** (712) STAT protein
4. Hs4507255 [KT] **KOG3667** (748) STAT protein
5. Hs6274552 [KT] **KOG3667** (750) STAT protein
6. HsM4507253 [KT] **KOG3667** (770) STAT protein
7. Hs21618338 [KT] **KOG3667** (769) STAT protein
8. Hs21618340 [KT] **KOG3667** (770) STAT protein
9. Hs21618342 [KT] **KOG3667** (794) STAT protein
10. Hs21618344 [KT] **KOG3667** (787) STAT protein

10 nodes with Node Degree: 9 = min, 9 = max, 9.00 = avg

Publications & Presentations

- Consistent Triplets of Candidate Paralogs by Graph Clustering International Joint Conference of InCoB, AASBi and KSBI (**BIOINFO** 2005) **Korea, 2005**
- Protein Domain Extraction by Quasi-convex set functions The Eleventh International Conference on Intelligent Systems for Molecular Biology (**ISMB**) **Sydney, 2003**
- Multi-alignment of Paralogs for Functional Annotation: Application to the Rice Genome. Proc. of the Fifth Annual Conference on **Computational Genomics. Baltimore, 2001**

Conclusions

- Consistent triplets: structural consistency.
- Fast and robust: no multi-alignment required.
- Criterion uses the minimum number of consistent triplets per node, which maximized for a cluster.
- Fully automatic without any help from experts.
- Coherence with KEGG: above 90% with better structural consistency of cluster classification.
- Almost all COG/KOG clusters are identified.
- New clusters are found from COG.
- Results can be viewed on the web at **http://www.cs.rutgers.edu/~seabee/**

Future Work

- CT-clustering can be applied for clustering structural objects. One application of CT-clustering can be to **images**:

1. Assign every informative pixel to some specific *features* in an image, like the sequence position is characterized by the related amino acid.

2. Find a matching method which constructs a correspondence between any two sets of informative pixels related to the corresponding two images like we used an alignment between two protein sequences.

Acknowledgement

- Dr. Casimir Kulikowski
- Dr. Ilya Muchnik
- Dr. Craig Nevill-Manning
- Dr. Joseph Mullat
- Dr. SukMoon Chang
- Dr. Dmitriy Fradkin, Dr. Luo
- Akshay Vashist, Thang Le
- Andrei Anghelescu, Jiankuan Ye