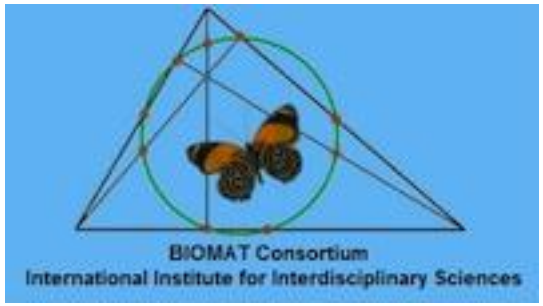



13th BIOMAT

INTERNATIONAL SYMPOSIUM ON MATHEMATICAL AND COMPUTATIONAL BIOLOGY


16th Oral Session

MINING THE CONSTRAINTS OF PROTEIN EVOLUTION

Fernando Encinas Ponce
Antonio Basilio de Miranda

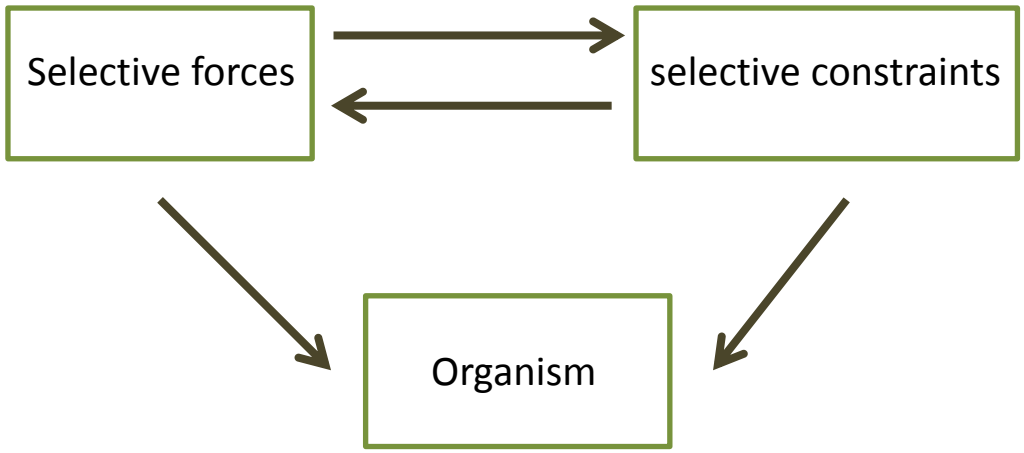

November, 6th, 2013 – Toronto, Canada

# OUTLINE

- Introduction
- Motivation
- Systems biology approach
- Computational methods - Data Mining
- Methodology
- Results
- Conclusions

Evolution

accumulation of random changes in the genetic makeup over time



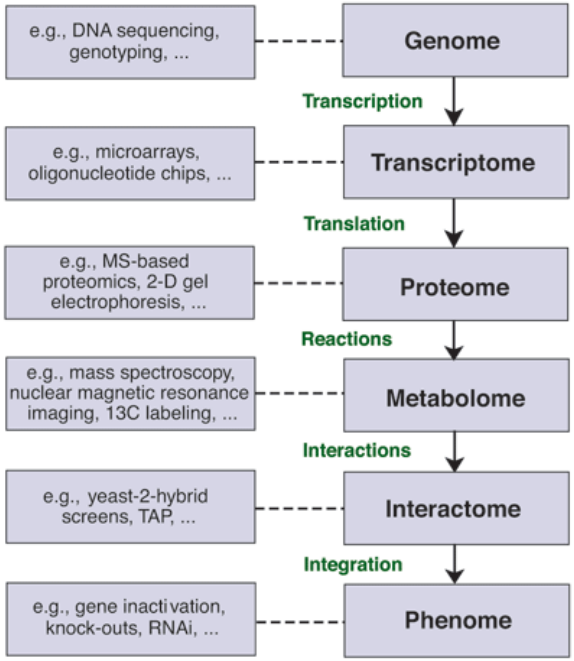

Evolutionary constraint: restriction or limitation on the course or outcome of evolution

Why proteins evolve at differente rates?

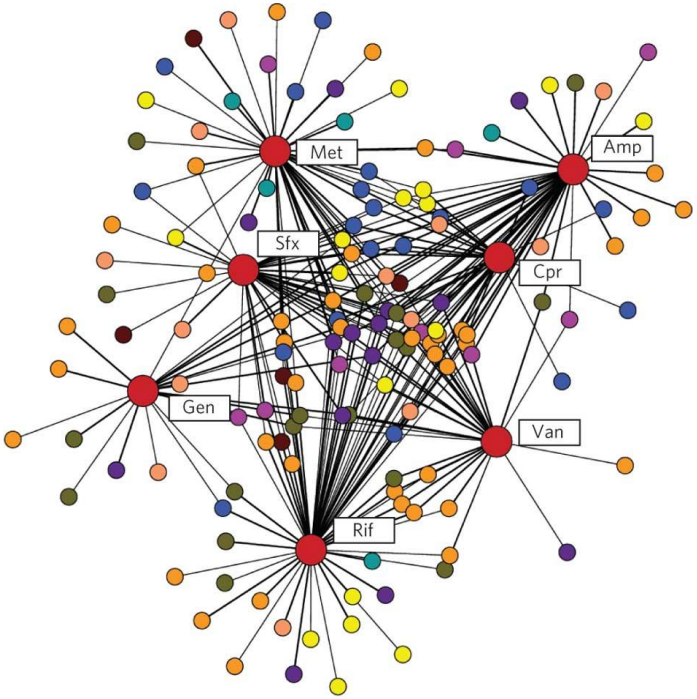
The pre-genomic view…

- Structure
- Fuction

# High-throughput biological information



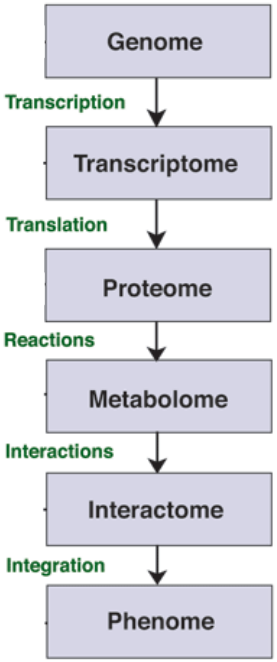
Biological complexity:




| Functional classification | |
|---|---|
| DNA binding, replication, recombination and repair | |
| Gluthatione biosynthesis and redox homeostasis | |
| Transport, efflux, cell-wall and cell-membrane synthesis | |
| Chaperones and proteases | |
| Protein synthesis | |
| General metabolic reactions | |
| Regulation | |
| Prophage-encoded genes and cell adhesion | |
| Unassigned genes | |

Antibiotics as probes of biological complexity
Shannon B Falconer, Tomasz L Czarny & Eric D Brown
Nature Chemical Biology 7,415–423 (2011)

among these data…

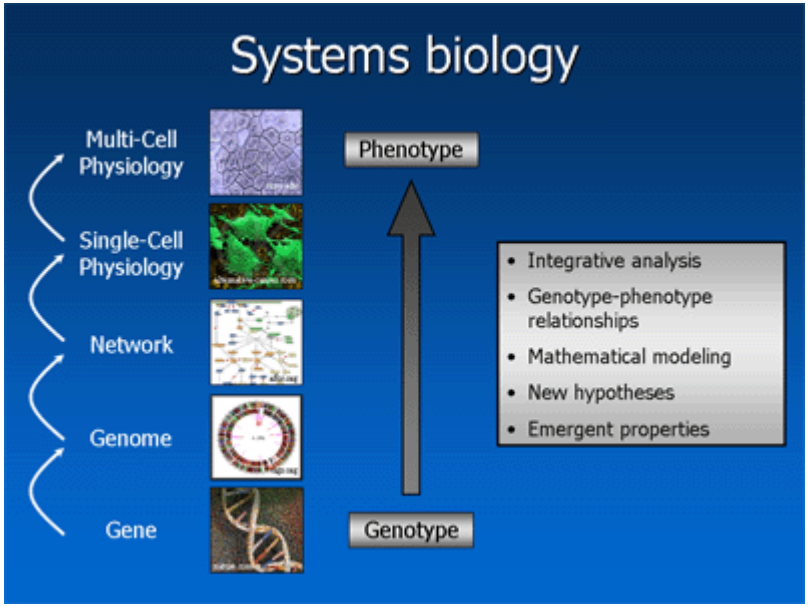



*Selective constraints?*

# MOTIVATION

- More and potential selective constraints may exist
- Natural selection works to optimize organisms

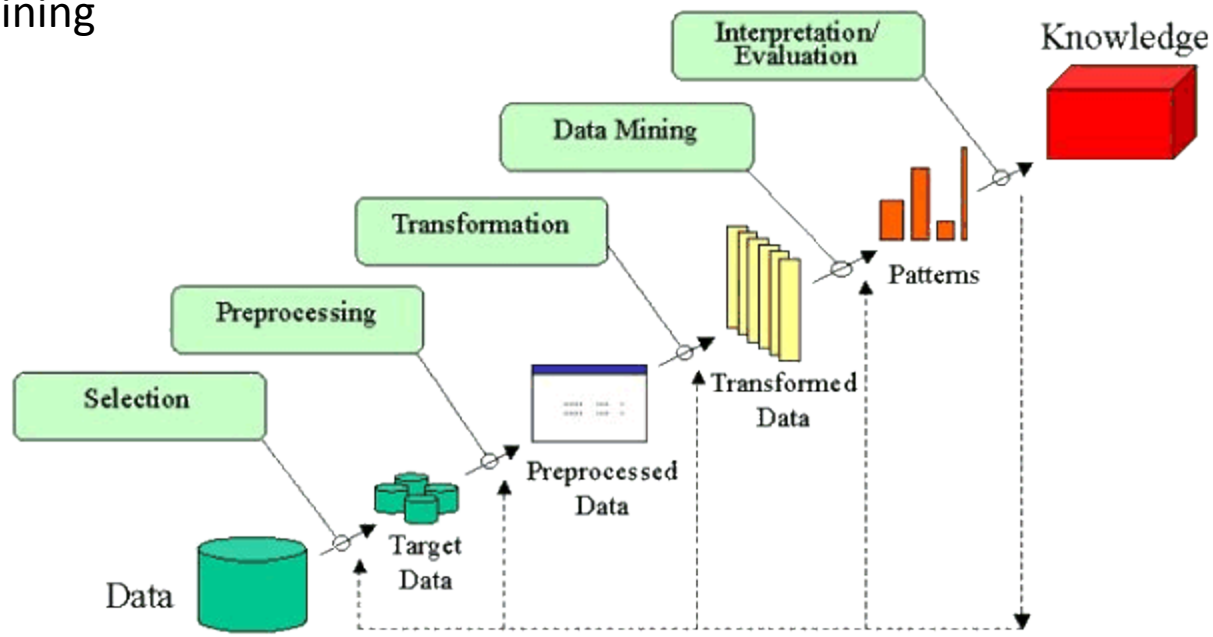
# SYSTEM APPROACH




Holistic perspective…

http://bme.virginia.edu/csbl/about.php

Data Mining




http://www2.cs.uregina.ca/~dbd/cs831/notes/kdd/1_kdd.html

# OBJECTIVES

- To identify overlooked genomic constraints that govern protein evolution
- To integrate information on these constraints into a single framework (biological system)

# METHODOLOGY


C

R

terms

evolutionary rates

MEGA

orthologs

Perl

CDS

gene data set

Databases

Hierarchical clustering of variables

Multiple Factor Analysis

Bayesian Factor Analysis

R

**Text mining**

**Data collection**

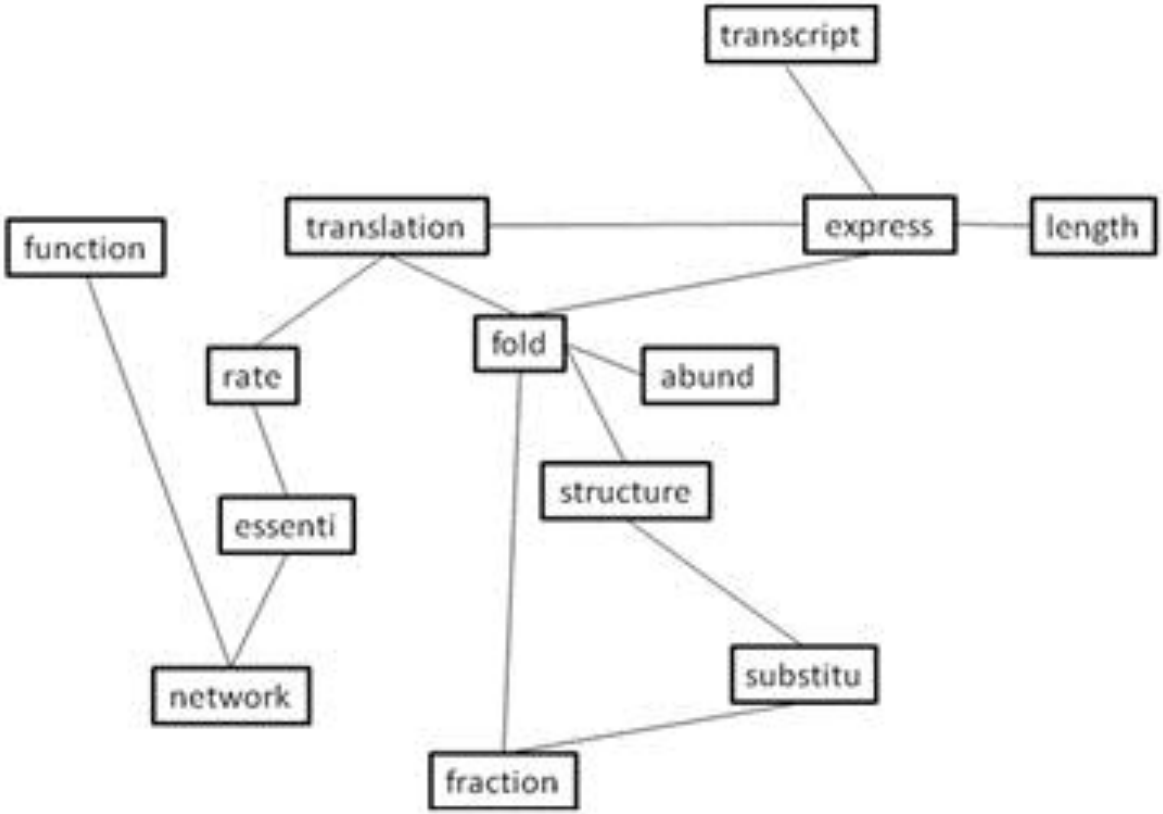
**Data analysis**

Text mining

**List of most frequent terms in the collection of documents**

| | | |
|---|---|---|
| [1] "chang" | "correl" | "data" |
| [4] "differ" | "effect" | "evolut" |
| [7] "evolutionari“ | "evolv" | "express" |
| [10] "figur" | "function" | "gene" |
| [13] "genom" | "interact" | "level" |
| [16] "mutat" | "network" | "ortholog" |
| [19] "protein" | "rate" | "relat" |
| [22] "residu" | "result" | "select" |
| [25] "sequenc" | "site" | "speci" |
| [28] "structur" | "studi" | "use" |
| [31] "yeast | | |

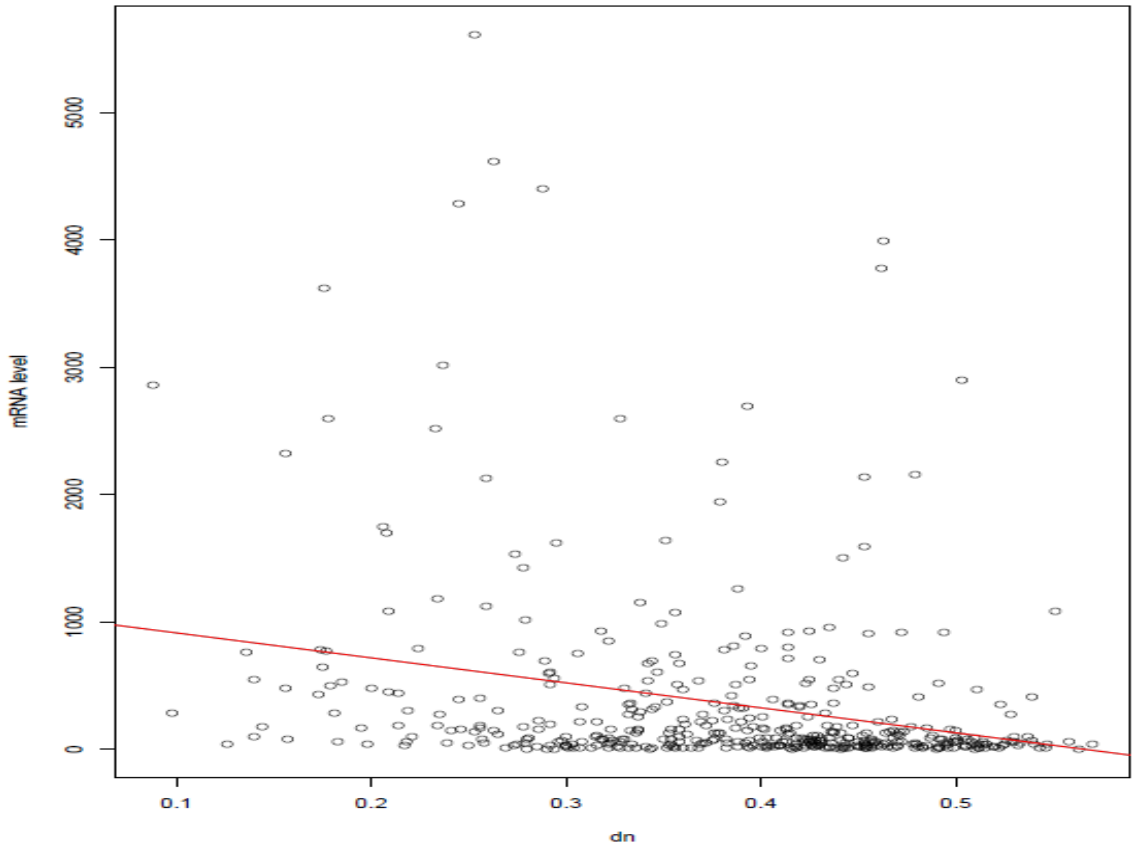
Term association analysis



## Text-derived variables

| Stem term | Gene/protein feature | Variable type | Nature |
|---|---|---|---|
| rate, substitu | number of synonymous substitutions (dS) | continuous | Evolutive |
| rate, substitu | number of non-synonymous substitutions(dN) | continuous | Evolutive |
| express | mRNA level | continuous | Expression |
| abund | protein level | continuous | Expression |
| translation | translation efficiency | continuous | Expression |
| length | protein length | continuous | Structural |
| structure | native structure | categorical | Structural |
| structure | instability index | continuous | Structural |
| structure | Stability | categorical | Structural |
| network | number of interactions | continuous | Functional |
| region, structure | low complexity porcentage | continuous | Structural |
| essenti | Essentiality | categorical | Functional |
| essenti | Dispensability | continuous | Functional |

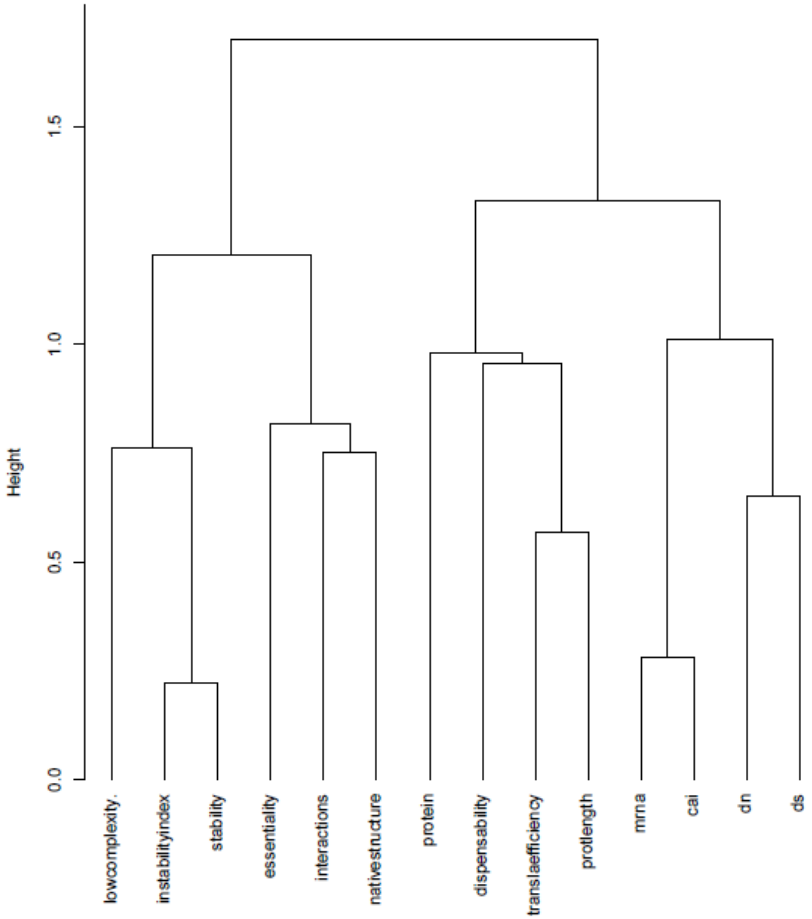
Pair-wise correlation analyses




Negative correlation between mRNA level and dN

| Variable | ORF length | dN | dS | dN/dS | mRNA level | Translational efficiency | Protein abundance | CAI | Number of interactions | Dispensability | % Low complexity | Protein length | Instability index |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ORF length | 100 | 17.74 | -19.12 | 14.13 | 3.17 | -51.78 | 6.31 | -12.03 | 3.69 | 8.25 | 2.54 | 99.35 | 8.33 |
| dN | | 100 | -41.15 | 62.6 | -26.68 | -25.06 | 3.49 | -49.6 | -6.77 | 1.42 | 9.16 | 17.54 | 22.22 |
| dS | | | 100 | -89.4 | -5.13 | 10.65 | 2.32 | -1.86 | -3.11 | -3.27 | 1.54 | -20.14 | 3.17 |
| dN/dS | | | | 100 | -9.77 | -13.4 | -0.74 | -19.56 | 0.37 | 2.32 | 2.16 | 15.35 | 4.91 |
| mRNA level | | | | | 100 | 8.15 | -2.84 | 71.14 | -3.53 | -2.47 | 3.73 | 3.25 | -17.54 |
| Translational efficiency | | | | | | 100 | -11.63 | 29.92 | 0.1 | -9.57 | -10.38 | -51.33 | -18.18 |
| Protein abundance | | | | | | | 100 | -7.38 | -2.01 | -2.06 | 1.64 | 6.32 | 9.66 |
| CAI | | | | | | | | 100 | -5.2 | -4.14 | 5.78 | -11.67 | -27.04 |
| Number of interactions | | | | | | | | | 100 | -1.25 | 9.52 | 3.7 | 9.93 |
| Dispensability | | | | | | | | | | 100 | -4.44 | 8.58 | -5.06 |
| % Low complexity | | | | | | | | | | | 100 | 2.66 | 38.93 |
| Protein length | | | | | | | | | | | | 100 | 7.89 |
| Instability index | | | | | | | | | | | | | 100 |

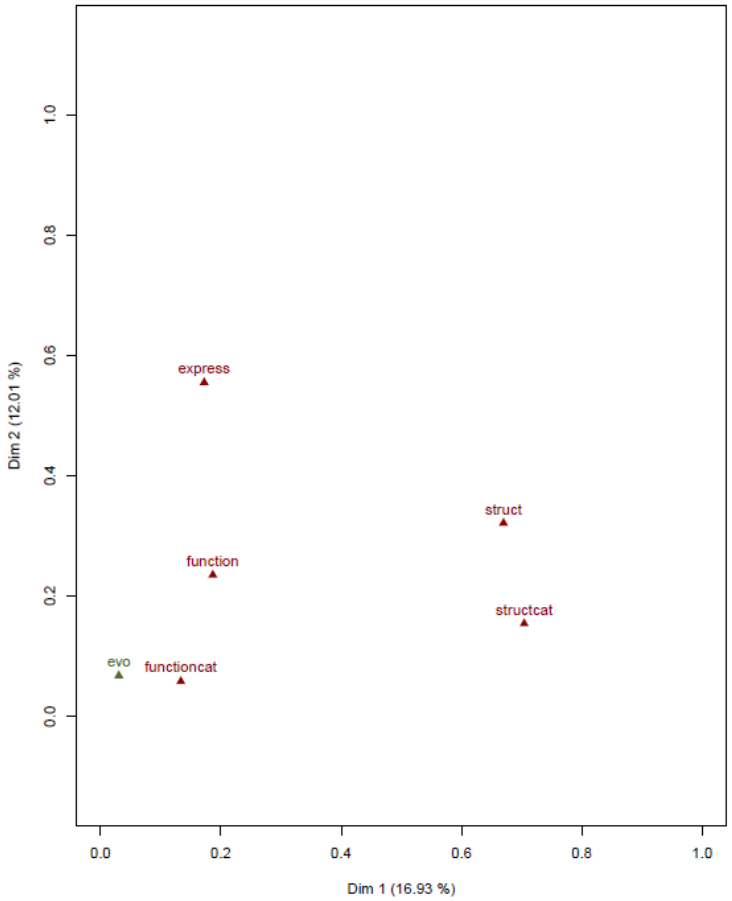
Clustering of variables reveals the underlying structure of the data



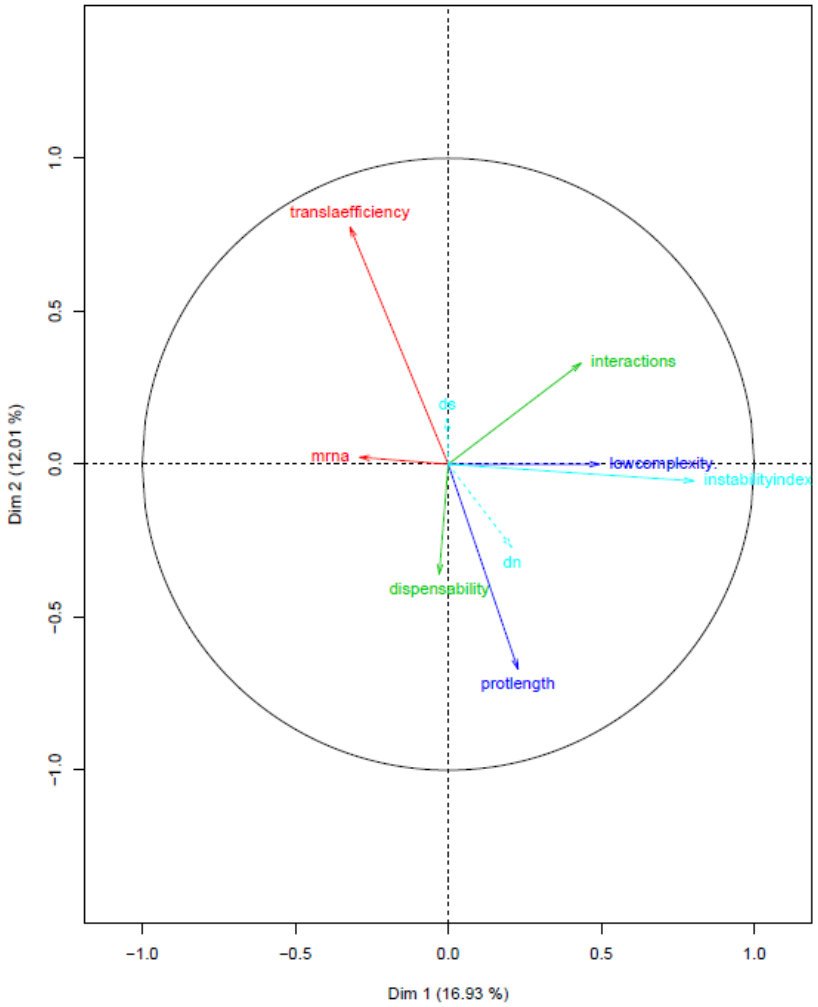
# Muliple Factor Analysis

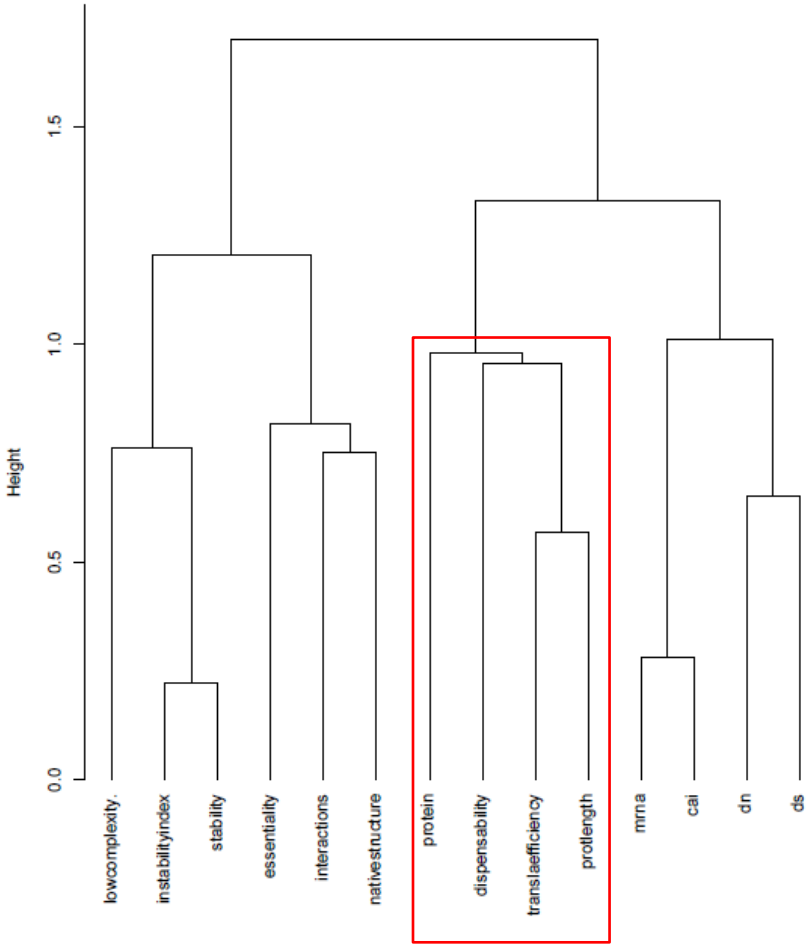
| Stem term | Gene/protein feature | Variable type | Nature |
|---|---|---|---|
| rate, substitu | number of synonymous substitutions (dS) | continuous | Evolutive |
| rate, substitu | number of non-synonymous substitutions(dN) | continuous | Evolutive |
| express | mRNA level | continuous | Expression |
| abund | protein level | continuous | Expression |
| translation | translation efficiency | continuous | Expression |
| length | protein length | continuous | Structural |
| structure | native structure | categorical | Structural |
| structure | instability index | continuous | Structural |
| structure | Stability | categorical | Structural |
| network | number of interactions | continuous | Functional |
| region, structure | low complexity porcentage | continuous | Structural |
| essenti | Essentiality | categorical | Functional |
| essenti | Dispensability | continuous | Functional |

# Latent constructs are useful to integrate genomic data



Individual coordinates show relationships between variables




Height
0.0
0.5
1.0
1.5
lowcomplexity.
instabilityindex
stability
essentiality
interactions
nativestructure
protein
dispensability
translaefficiency
protlength
mrna
cai
dn
ds

## Bayesian Factor Analysis

Positive and negative contributors to an adapted protein translation system

| | Factor loading | Psi-uniqueness |
|---|---|---|
| synonymous substitutions | 0.4121 | 0.6921 |
| instability index | -0.213 | 0.9548 |
| translation efficiency | 0.8783 | 0.2129 |
| protein level | -0.1410 | 0.9826 |
| Dispensability | -0.0995 | 0.9954 |

# Convergence diagnostic test

| | Stationary | p-value |
|---|---|---|
| dS | Passed | 0.243 |
| instability index | Passed | 0.180 |
| translation efficiency | Passed | 0.122 |
| protein level | Passed | 0.165 |
| dispensability | Passed | 0.584 |
| Psi-dS | Passed | 0.608 |
| Psi-instability index | Passed | 0.219 |
| Psi-translation efficiency | Passed | 0.104 |
| Psi-protein level | Passed | 0.380 |
| Psi-dispensability | Passed | 0.454 |

# CONCLUSIONS

- Innovative computational methods are needed to make sense of biological data
- Translational efficiency, structural instability and low complexity regions appear to be important determinants of protein evolution
- Latent constructs are an interesting aternative to integrate genomic data.

Thank you!!!