Protein Spaces

Misha Gromov

September 26, 2019

This is an annotated extract from my 2019 course at CIMS called "Mathematics of Life Spaces"

1 Three Perspective on Proteins

There is a three way fork in the road toward mathematical protein spaces.

I. *Individual Protein Molecules: Folding, Structure Function.* Start with pictorial images of proteins such as these.

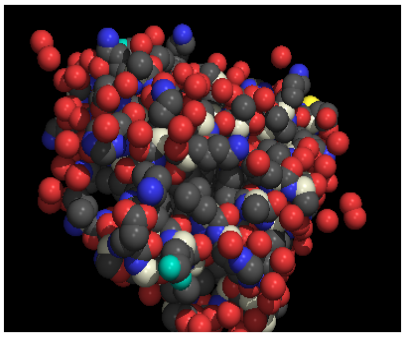

Figure 1: Spatial model of a folded protein molecule

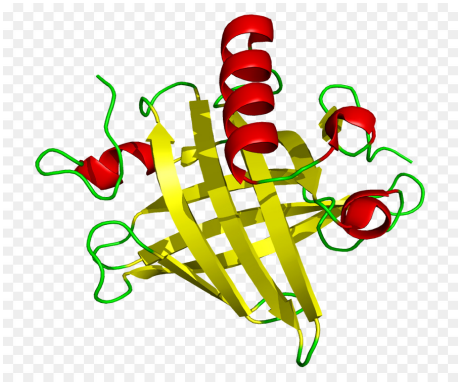

Figure 2: Schematic protein structure.

Then look at the pictures of proteins arriving to these shapes by the folding process.

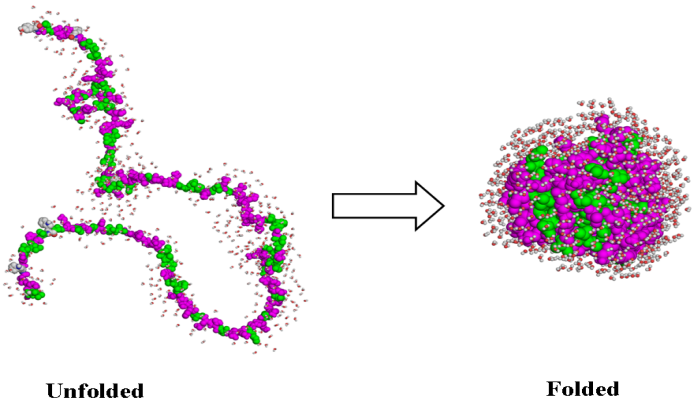

Figure 3: Protein folding in water: little dots depict water molecules which surround proteins and which are pushed out of the interior regions of the folded protein molecules.

Making math out of this is hard: following in steps after P.J. Flory and Orr,[1] we turn to something easy: unfolded proteins, thought of as chains of beads freely floating in solution.

But is it easy? Not at a all:the space of these looks something like this.

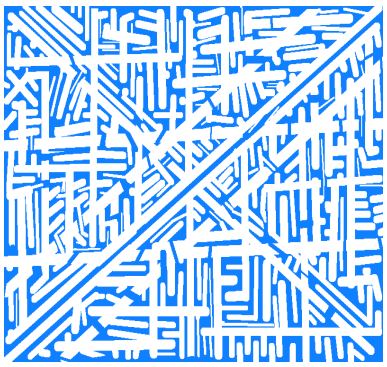

Figure 4: The square represents the space of all chains of points $x_1, ..., x_i, ..., x_N$ in the 3-space, such that $dist(x_i, x_{i+1}) = 1$. The white strips correspond to the positions of the pairs (x_i, x_j), where $dist(x_i, x_{j1}) \leq 2\varepsilon$, which corresponds to impossible positions of material ε-beads with the centers x_i and x_j. The remaining "blue" depicts the configuration space of chains of ε-beads in the 3-space.

Despite multiple mathematical papers dedicated to (discretized versions of) these "strings of beads in the space," called *self avoiding random walk(s)*, amaz-

[1]P.J. Flory, The configuration of a real polymer chain, J. Chem. Phys.17(1949), W.J.C. Orr. Statistical treatment of polymer solutions at in
infinite dilution. Transactions of the Faraday Society, vol. 43, (1947).

ingly little has been rigorously proved so far. But even if we prove 200% of what we want, it will contribute something like 0.0002% to our knowledge of proteins.

The next level of approximation to "reality" after the self avoiding random walk model is that of *the energy landscape of a protein.* This is a real valued function E on the space $\mathcal{S}$ of self avoiding walks.

An essential characteristic of this function is the *weighted Kronrod (Adelsson-Veleski-Reeb-Stein) tree* $\mathcal{T} = \mathcal{T}_E$ that is the set of connected components of the sublevels

$$\{S \in \mathcal{S}\}_{E(S)\leq h, h\in\mathbb{R}}.$$

The function $E : \mathcal{S} \to \mathbb{R}$ obviously factorizes via $\mathcal{T}$,

$$\mathcal{S} \to \mathcal{T} \to \mathbb{R},$$

where the levels of the quotient map $\mathcal{S} \to \mathcal{T}$ are connected and all continuous maps $\mathcal{S} \to \mathcal{T}'$ with connected pullbacks canonically factor as $\mathcal{S} \to \mathcal{T}' \to \mathcal{T}$.

The weight we speak of is a measure on $\mathcal{T}$ that is the push forward of the natural measure/volume on $\mathcal{S}$.

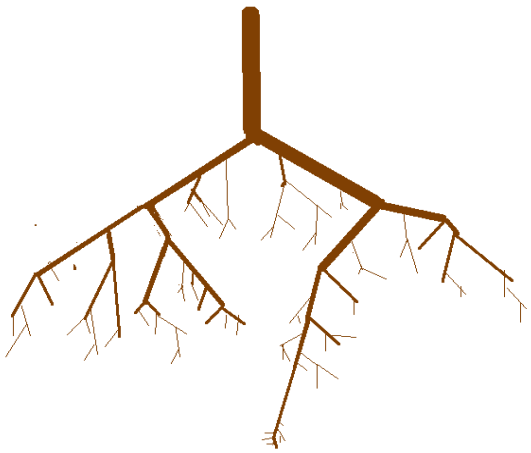

Figure 5: The energy is represented by the height function and the weight by the thickness of the branches of the tree. The leaves correspond to the local minima of the energy with the global minimum positioned at the lowest leaf; this corresponds to the folded protein molecule

The energy of the folded protein S, e.g the one serving an enzymatic/catalytic function may be not fully localized, but distributed over branches of a smaller tree where chemical reaction(s) between molecules can be seen as certain operations over such trees.

The simplest such operation is the *convolution of trees*, call them $\mathcal{T}_1$ and $\mathcal{T}_2$ with energies $E_i : \mathcal{T}_i \to \mathbb{R}$ and weights/measures on μ_i and on them, $i = 1, 2$. This convolution, which corresponds molecules which don't interact is defined as

$$\mathcal{T}_1 * \mathcal{T}_2 = \mathcal{T}_{E_1+E_2},$$

where $E_1 + E_2$ is the energy function on the product $\mathscr{T} = \mathcal{T}_1 \times \mathcal{T}_2$ with the measure $\mu_1 \otimes \mu_2$, where the (2-dimensional) space $\mathscr{T}$ plays the role of $\mathbb{S}$ in the definition of the Kronrod tree.

It would be amusing to reformulate what is known (and unknown) about enzymatic catalysis in the multidimensional geometric language[2] of this kind, which may be more instructive than how it is depicted in the traditional way exemplified by figure 6 below.

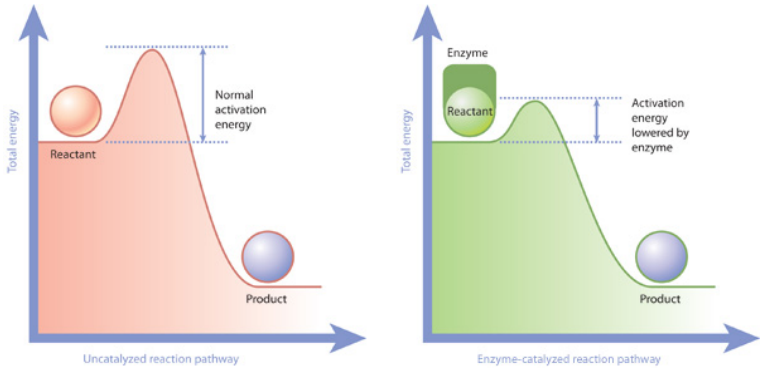

Figure 6: Enzyme speeds up a reaction by lowering the activation energy barrier

II. *Sequence spaces, Proteins Families Evolution trees.* Proteins P are (almost) fully[3] defined by what is called their primary structure that is a sequence of 20 letters for 20 basis amino acids and thus, can be thought as points in the product $\{20A\}^*$ where $\{20A\}$ denotes the stands for the set of 20 amino acids and $*$ for a moderate number N, something between 50 and 300.

Sequence spaces such as $\{20A\}^300$, especially $\{0,1\}^N$ have been extensively studied and are as much understood, as anything else in mathematics.

In particular, one has a clear idea of the geometry of the *Hamming metric* between sequences $(a_1, ...a_N)$ and $(b_1, ..., b_N)$ that is the number of positions $i \in \{1, ...N\}$, where $a_i \neq b_i$, that is, biologically speaking, the (minimal) number of *substitution point mutations* needed to pass from $(a_1, ...a_N)$ and $(b_1, ..., b_N)$.

What is harder is to understand the properties of the subset $\mathcal{P} \subset \{20A\}^*$ of sequences corresponding to actual proteins in living organisms (that us similar to the set of sequences in 26 letters in the sentences ever written anywhere in English).

These constitute a minority of all 300-long sequences, say $< 10^{25}$ out of $20^{300} > 10^{390}$ could have been be present among proteins in organisms who have ever lived on Earth about with about 100 000 000 recorded sequences recorded today[4] the essential structure in subset $\mathcal{P} \subset \{20A\}^*$ is that of *directed graph*, where the arrow $P \to P'$ signifies a mutation, e.g. the above point substitution,

[2]Dimension may play an essential role, since participation of an enzyme in a reaction increases the dimension of the parameter space of this reaction and which may be relevant for describing the mechanism which lowers the activation (free) energy of the reaction, where the increase of dimensionality may radically modify not only and not so much the energy as the "entropic geometry" of the process. (Think of the entropy of the lock/key system.)

[3]Proteins may undergo *post-translational modifications* not encoded by their amino acid sequences.

[4]For comparison, the number of atoms on Earth is $\approx 10^{50}$.

where an amino acid a_i is substituted by a'_i.[5]

Then the protein sequences are seen as tree-like clusters in the full sequence space $\mathcal{P} \subset \{20A\}^*$, where, in fact, only the leaves of these are visible in the today proteins, while the rest of the tree is obtained by the (conjectural) *ancestral sequence reconstruction*, (mostly) based on the *multiple sequence alignment* algorithms.

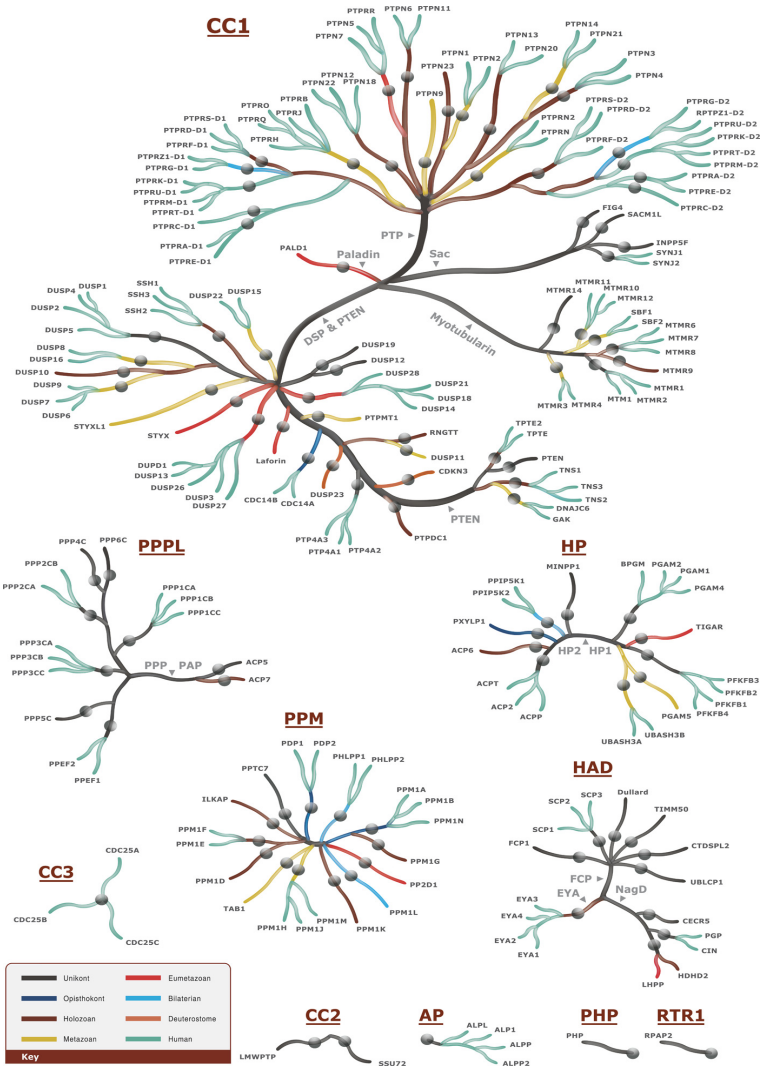

Figure 7: Evolution trees of protein phosphatases

As far as the functions of proteins go, and this is what, besides neutral mutations, drives the evolution, the shape of folded proteins is more essential than their sequential structure.

[5]Other kinds of mutations are discussed later on

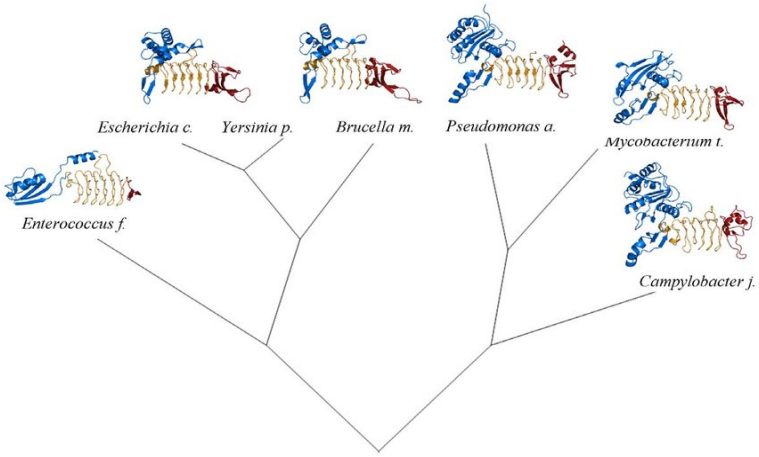

Figure 8: Evolutionary tree of DapD enzymes in bacteria

But there in no(?) apparent formal mathematical approach to ancestral reconstruction of protein folds.

III. *Proteins Interaction Networks.* The two basic classes of relations between Life's childrens, be they macromolecules or sentences in a language, are

similarity and cofunctionality.

The main source of similarity in biology is common ancestry, while the most apparent manifistation of cofunctonality of two proteins in the cell is their tendency to preferentially bind one to another.

Here is how the resulting graph(s) may look like,

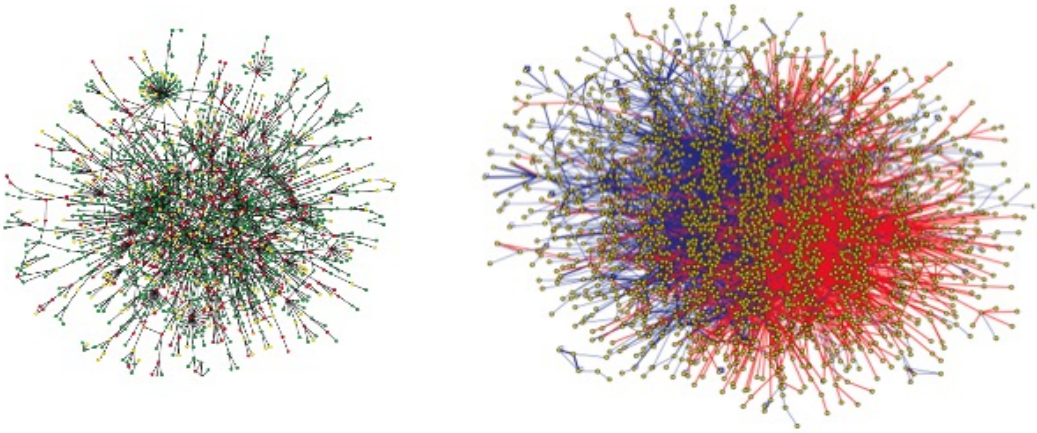

Figure 9: Yeast (left) and human (right) interactomes obtained using the yeast-two hybrid method.

and below is a fragment of such a graph.

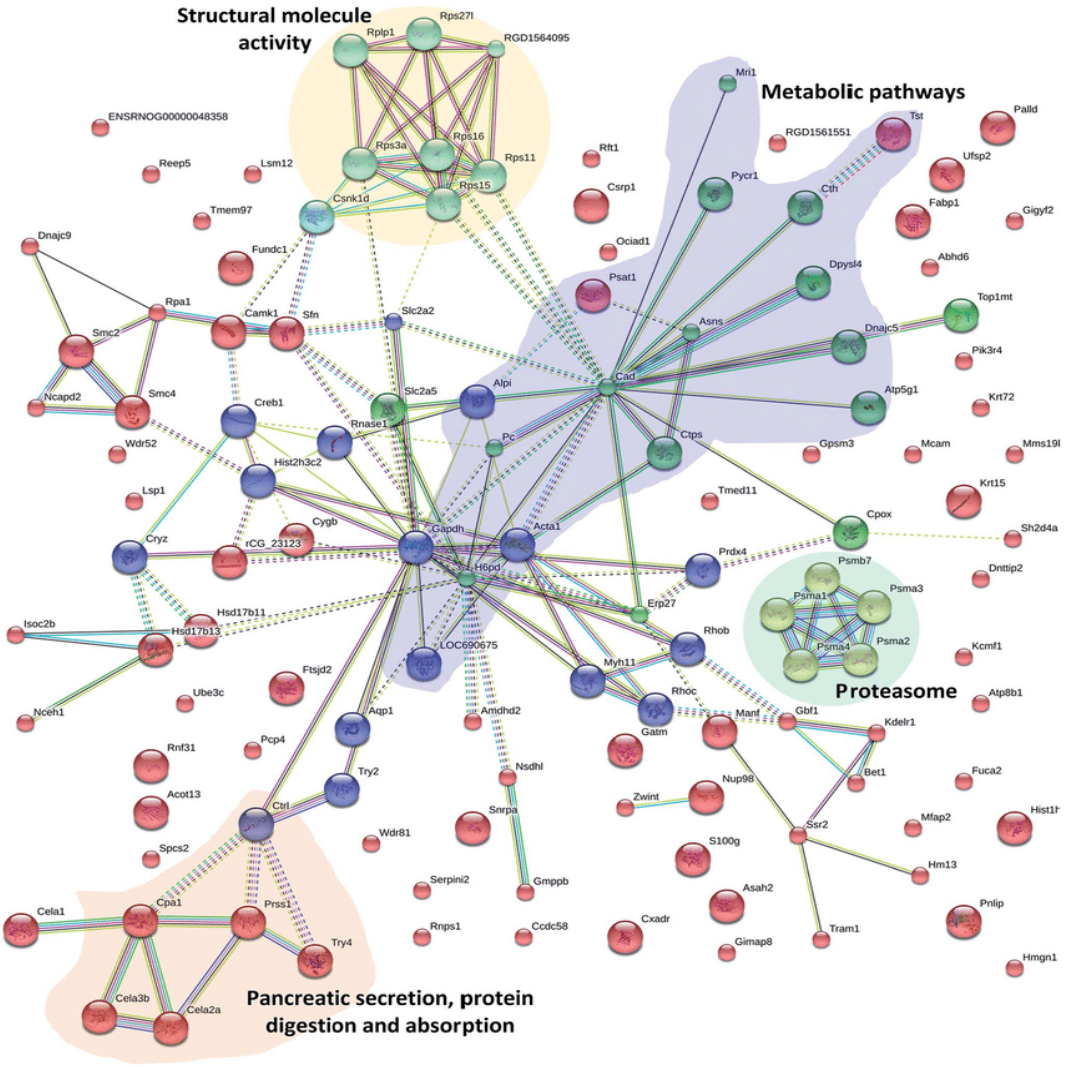

Figure 10: Protein-protein interaction analysis of cluster 4 highlighting protein networks involved in pancreatic secretion, protein digestion and absorption (orange), metabolic pathways (purple), proteasome (green) and structural molecule activity (yellow). In the network, proteins are represented as nodes. Colors of the lines connecting the nodes represent different evidence types for protein linkage.

The above **I,II,III** reperesent only an outline of the architecture of the Protein Unoverse, where, observe, the ancestrial similarity of **II** applies to the conbiantorics depicted by **III**.

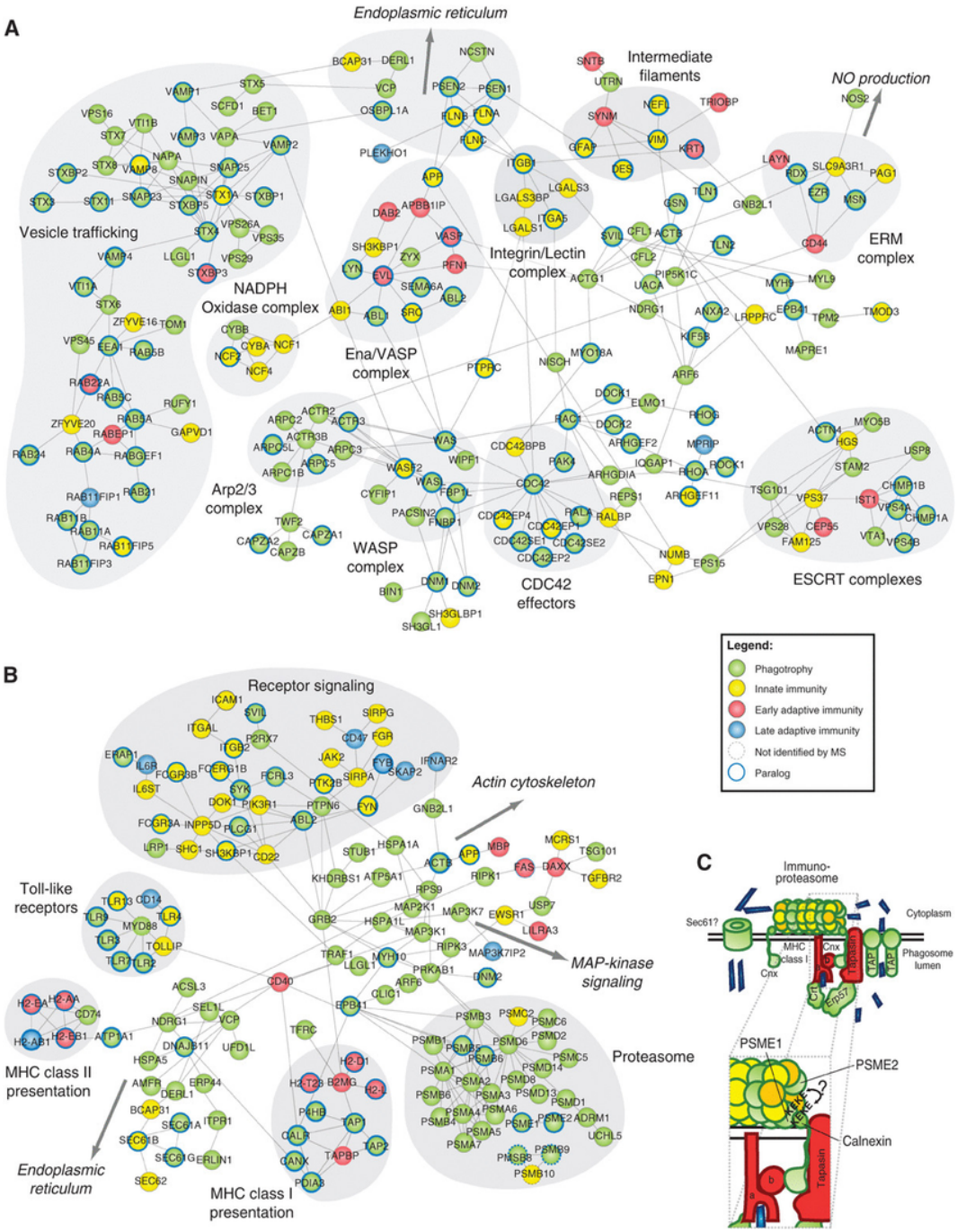

Figure 11: Evolution of the phagosome proteins network

There are (at least) two ways to treat this mathematically.

1. Developing a mathematical teory of random network evolution.

2. Reconstructing ancesral networks by kind of multiple "alignment" of extant combinatorial network structures.