Phylogeny and Sequence Space: A Combined Approach to Analyze the Evolutionary Trajectories of Homologous Proteins. The Case Study of Aminodeoxychorismate Synthase

Sylvain Lespinats, Olivier de Clerck, Benoît Colange, Vera Gorelova, Delphine Grando, Eric Maréchal, Dominique van Der Straeten, Fabrice Rébeillé, Olivier Bastien

TITLE:

Phylogeny and sequence space: a combined approach to analyze the evolutionary trajectories of homologous proteins. The case study of aminodeoxychorismate synthase

Authors:

Sylvain Lespinats[1], Olivier De Clerck[2], Benoît Colange[1], Vera Gorelova[3,4], Delphine Grando[5], Eric Maréchal[5], Dominique Van Der Straeten[3], Fabrice Rébeillé[5] and Olivier Bastien[5,*]

Adresses:

(1) Univ. Grenoble Alpes, INES, F-73375, Le Bourget du Lac, France

(2) Department of Biology, Phycology Research Group, Ghent University, Krijgslaan 281, 9000 Gent, Belgium;

(3) Department of Biology, Laboratory of Functional Plant Biology, Ghent University, K.L Ledeganckstraat 35, 9000 Gent, Belgium;

(4) Department of Botany and Plant Biology, Laboratory of Plant Biochemistry and Physiology, University of Geneva, Quai E. Ansermet 30, 1211-Geneva, Switzerland.

(5) *Univ. Grenoble Alpes, CEA, CNRS, INRA, BIG-LPCV, 38000 Grenoble, France*

Corresponding author:

* Olivier Bastien

UMR 5168 CNRS-CEA-INRA-Université J. Fourier, Laboratoire de Physiologie Cellulaire Végétale, Département Réponse et Dynamique Cellulaire, CEA Grenoble, 17 rue des Martyrs, F-38054, Grenoble cedex 09, France

Tel : +33 (0)4 38 78 49 85

E-mail : olivier.bastien@cea.fr

Keywords

Phylogenetic method, Multi-Dimensional Scaling, Sequence Space, ADC Synthase, Phylogenetic Tree

ABSTRACT

During the course of evolution, variations of a protein sequence is an ongoing phenomenon however limited by the need to maintain its structural and functional integrity. Deciphering the evolutionary path of a protein is thus of fundamental interest. With the development of new methods to visualize high dimension spaces and the improvement of phylogenetic analysis tools, it is possible to study the evolutionary trajectories of proteins in the sequence space. Using the Data-Driven High-Dimensional Scaling method, we show that it is possible to predict and represent potential evolutionary trajectories by representing phylogenetic trees into a 3D projection of the sequence space. With the case of the aminodeoxychorismate synthase, an enzyme involved in folate synthesis, we show that this representation raises interesting questions about the complexity of the evolution of a given biological function, in particular concerning its capacity to explore the sequence space.

Contact: olivier.bastien@cea.fr

1. INTRODUCTION

The determination of the main parameters that can explain the evolutionary dynamics (velocity and trajectory) of a protein in a mathematical abstract sequence space remains a major difficulty in the study of molecular evolution (Starr and Thornton 2016). For more than 40 years, studies of protein evolution focused on the localization of the amino acid substitutions, their rate of appearance and the phylogenetical relationships between the considered protein sequences (Dayhoff 1976; Dayhoff et al. 1983; Lemey et al. 2009). More recently, the study of evolutionary trajectories and the capacity to determine a fitness for a protein has gained a lot of interest and accuracy because of the large amount of available sequence data and the development of new experimental techniques such as directed evolution (Kondrashov and Kondrashov 2015; Romero and Arnold 2009) or deep mutational scanning (Starr and Thornton 2016). Assuming that better adapted organisms will reproduce faster and will more spread their genes throughout the population (Wright 1931), we can define the "fitness" for a given protein as a measure associated with the host organism's ability to produce a functional protein sequence in a given environment (Romero and Arnold 2009). The key concept for understanding the interplay between a protein sequence, its physical or biological properties, its fitness and its evolution is the sequence space. A sequence space is a configuration space. One of the very first definitions of the sequence space was proposed by Maynard Smith (1970) who defined it as a multidimensional representation of all possible protein sequences (nodes), each connected to its closest neighbors by links (edges) representing changes at a unique residue position (Maynard Smith 1970). By assigning a physical or biological property (like fitness) to each sequence, we can define a scalar field (that is a function defined on a space whose value at each point is a scalar quantity, as for example temperature or pressure on the Earth surface), resulting in a "topographic map" of the sequence space. By analogy, this can be compared to a topographical map of a geographic landscape where elevations of locations are indicated for

each point specified by latitudinal and longitudinal coordinates, and wherein effects like epistasis on the fitness landscape can be studied. Thereby, "topographical" observations are expected to inform on the topological properties of the protein sequence space as well as on the dimension (named intrinsic dimension, Facco et al. 2017) and topological properties of the subspace where protein sequences are actually moving in (Facco et al. 2017). Protein evolution can then be understood as a walk on this high-dimensional fitness landscape, in which regions of higher elevation represent adapted/stable functional proteins and in which evolutionary paths are the "safe" tracks connecting these regions (Kondrashov and Kondrashov 2015, Starr and Thornton 2016).

Another common analogy is to compare the protein sequence with the Universe and the distribution of matter and energy into it. This leads to the concept of a 'protein universe', which contains the totality of all possible proteins (Holm and Sander 1996, Koonin et al. 2002). Using an Information theory approach (Shannon and Weaver 1949), Adami et al. (2000) have shown that the genomic complexity (and so the exploration of the sequence space) is forced to increase because of natural selection pressures within a fixed environment. In addition, changes in sequence length (leading to an expansion of the sequence space) provide new spaces (called by the author ''blank tape'') in which environmental information and increase of complexity associated with the ongoing evolution can be recorded (Adami et al. 2000). Thus, by continuously feeding databases with novel sequences, the topographical map extends, allowing the exploration of still unsuspected fitness landscapes. With all these considerations, the size (dimension) of the protein universe (*i.e.* the total number of possible protein sequences) can be viewed as infinite (Adami et al. 2000). Actually, with an average protein length between 283 for Archaea and 438 for Eukaryotes (Lukasz and Kozlowski 2017), the number of different protein sequences can be estimated between 20^{283} and 20^{438} (i.e. between 10^{556} and 10^{866}). By comparison, this number is far much greater than the number of atoms in our physical Universe

which is "only" 10^{79}-10^{80}. However, only a small fraction of the protein sequence space is populated by real protein sequences. Indeed, the number of unique sequences encoded in the actual genomes is substantial and can be estimated to be nearly 5.10^{10} (Koonin et al. 2002). Obviously, the repertoire of acceptable amino-acid substitutions and hence the potential motion in the sequence space is severely restricted by natural selection so that both structural and functional integrity of evolving proteins are maintained (DePristo et al. 2005; Tokuriki and Tawfik 2009). A protein sequence is not only a linear string of amino acids, it is also self-organized in secondary structures including alpha-helices and beta-sheets that contribute to form and stabilize a three-dimensional conformation. The general organization of alpha-helices, beta-sheets, coil domains *etc*., are referred to as a "fold". From this point of view, it was shown that the number of possible folds is finite (nearly 10,000 folds), which means that the sequence space occupied by protein folds contains "hot spots", where most sequences are gathered, whereas the rest is almost empty (Koonin et al. 2002). Interestingly, the distribution of size of a fold region follows asymptotic power laws, typically associated with scale-free networks and scale invariance of the studied object, suggesting that genome evolution is driven by extremely general mechanisms based on the preferential attachment principle (Koonin et al. 2002).

Estimating upper and lower limits for the number of organisms, genome sizes, mutation rates and for the number of functionally distinct classes of amino acids, Dryden et al. (2008) suggested that all the protein sequence space has already been explored by mutations and other genome editing processes since Life emerged 4 billions of years ago. By contrast, based on a computational approach, Povolotskaya and Kondrashov (2010) showed that ancient proteins are still diverging from each other, indicating an ongoing expansion of the protein sequence universe and that $3{,}53.10^{09}$ years have not been enough to reach the limit of this divergent evolution. These different conclusions could be explained by the fact that Povolotskaya and

Kondrashov (2010) focused not only on the number of possible functional sequences of a given length, but also on the rate of divergence of distant protein sequences.

All these considerations show that the questions of how sequences are distributed in the sequence space and how these sequences are moving in this space are of a fundamental and practical interest. Recently, we proposed a new way to explore these fundamental questions by mapping phylogenetic trees (which can be seen as evolutionary paths between sequences) onto a representation (*i.e.* an embedding) of the protein sequence space, and we applied this approach to explore the evolution of two enzymes, the dihydrofolate reductase (DHFR) and the dihydropteroate synthase (DHPS) involved in the biosynthesis of folates (Gorelova et al., 2019). Folate derivatives are central cofactors of the C1 metabolism, directly involved in the synthesis of nucleic acids, some amino acids (glycine, serine, methionine) and indirectly, through S-adenosyl methionine, in all methylation reactions (Rébeillé et al. 2006). Here, we improved and further describe this method which is summarized in figure 1. An example of application is also provided with the case of the aminodeoxychorismate synthase (ADCS), another enzyme involved in folate synthesis (Gorelova et al. 2019; Rébeillé et al. 2006).

2. MATERIALS AND METHODS

2.1 Identification of putative orthologs and protein analysis

To identify putative orthologs of the chorismate binding domain of the ADCS folate pathway genes, full-length protein ADCS sequence from Arabidopsis thaliana (Open Reading Frame, ORF, Name: AT2G28880) were used as a query to run a BLASTp on the NCBI non-redundant sequence database (Pruitt et al. 2005), the JGI genome portal (Nordberg et al. 2014) or specific databases such as, plasmoDB (Aurrecoechea et al. 2009) and TAIR (Berardini et al. 2015) (table 1). Phylogenetic studies (section 2.2) of the bifunctional enzyme ADCS were performed with the largest domain in each case, *i.e.* the Chorismate Binding Domain which have been identified using Conserved Domain Search Service v3.15 61 Parameters: Expect Value

threshold=0.01, with an E-value cutoff less or equal to 1e-15 (Altschul et al. 1990). cDNA sequences for the identified orthologs were also retrieved. To infer both putative orthology and the domain architecture of the retrieved proteins, the selected protein sequences were analyzed for the presence of functional protein domains using Conserved Domain Search Service v3.15 (Marchler-Bauer et al. 2011) with an expect value threshold equal to 0.01 and a composition-based statistics adjustment.

2.2 Phylogenetic analysis

The alignment of the Full-length cDNA sequence Chorismate Binding Domains was carried out at the amino acid level using MAFFT (Katoh and Standley 2013). This alignment was then used to generate the corresponding nucleotide sequence alignment using TranslatorX (Abascal et al. 2010). During the process, alignments was curated using both the multiple alignment editing softwares ClustalX 2.0 (Larkin et al. 2007) and Jalview 2 (Waterhouse and al. 2009). The phylogenetic relationship was inferred using the maximum likelihood method PhyML (Guindon et al. 2010). The evolutionary model selection was done with MEGA 7 (Kumar et al. 2016) under the Bayesian information criterion (Neath and Cavanaugh, 2012). According to the model selection process, the best evolutionary was GTR+G+I (Nei and Kumar 2000). The maximum likelihood tree was evaluated with aLRT (Anisimova and Gascuel 2006), a non-parametric branch support based on a Shimodaira-Hasegawa-like procedure. Phylogenetic trees were visualized using FigTree (Rambaut 2018).

2.3 Sequences as points in a multidimensional space

Each biological sequence can be represented in a so-called protein sequence space, for example the Configuration Space of Homologous Proteins (CSHP) (Bastien et al. 2005; Bornberg-Bauer

and Chan 1999; Dryden et al. 2008). Many classical distances between sequences in this space have been proposed, considering it as a metric space including "edit", "hamming" and "phylogenetic" distances (Setubal and Meidanis 1997), some of them being evolutionary distances. The choice of a specific distance depends on what is upon scrutiny. Here, we focused on evolutionary process and distances between sequences were computed using the classical PAM250 matrix with the Phylip prodist program (Felsenstein 1981).

Such set of items may be represented as points in a Euclidean space showing same distances between them as demonstrated in (Young and Householder, 1938). In that framework, each point may be seen as the representation of a sequence and distances between them would "encode" or "embed" the corresponding biological differences. Please notice that:

- the dimension of the Euclidean space needed by Young and Househoder's theorem to unfold items is not set a priori and depends on the distance matrix's features.
- the meaning of each axis of the resulting space would not be obtained by construction.

2.3.1 Intrinsic dimension of CSHP

Even if an upper boundary for the possible hypothetical biological sequence can be set by the 20 possible amino-acids (if gaps are not allowed) to the power of the number of residues that could be present/substituted in sequences (Bastien et al. 2005; Dryden et al. 2008),estimating the theoretical dimension of the sequence space is not easy. In that goal, if one takes a given sequence $\boldsymbol{\sigma}$, it is possible to define a referential called $\boldsymbol{R}$ considering each position in sequence $\boldsymbol{\sigma}$ as an axis (a dimension in the $\boldsymbol{R}$ referential). One can therefore place a homologous sequence $\boldsymbol{\sigma'}$ in $\boldsymbol{R}$. Each coordinate *i* of $\boldsymbol{\sigma'}$ in $\boldsymbol{R}$ will be a continuous function of the mutual information between residues in $\boldsymbol{\sigma}$ and $\boldsymbol{\sigma'}$ at the homologous substituted position *i* (Bastien et al. 2005). For example, we can consider two assumed homologous sequence **A** and **B** with two homologous

amino acids **a** (in **A**) and **b** (in **B**) which will be aligned in a multiple alignment at a position **p**. Then, if we choose **A** as a referential ***R***, the **p**-coordinate of **B** in ***R*** will be a function of the similarity between **a** and **b**. Obviously, the dimension of the space into which real biological homologs to $\boldsymbol{\sigma}$ are evolving can be much lower than those of the space ***R*** generated by $\boldsymbol{\sigma}$, the space of potential, or random sequences because many regions of the CSHP will not provide functional proteins leading to very low fitness values on these regions (Yau et al. 2015). Moreover, the topological dimension of the subspace wherein protein sequences are actually moving on can be much lower than the crude large number of coordinates, a common feature in high dimensional problems (Facco et al. 2017). This question is analogous to the problem of finding the position of a single particle moving in ordinary Euclidean 3D-space. Its position is defined *a priori* by a vector (x, y, z). Nevertheless, a particle might be constrained to move on a specific manifold if, for example, it is attached by a rigid linkage, free to swing around an origin, and hence is constrained to lie on a sphere. In this case, it is said that its configuration space is the subset of coordinates in $\mathbb{R}^3$ that define points on the sphere S^2. In this case, one says that S^2 is the configuration space and that it has a dimension equal to 2 (Gignoux and Silvestre-Brac 2002). In practice, the dimension of observed CSHP would also be limited by the number of considered sequences. Indeed, a set of n items in a multidimentional space can be embedded in a n-1 dimension Euclidean space while perfectly preserving their mutual distances (if no linear dependences can be found between data, the results is a simplex, otherwise, the dimension of the space can be chosen lower). However, the practical dimension of CSHP computed from real data will probably be higher than 2 or 3, and consequently not understandable by human eye. For that reason, non-linear multidimensional scaling method may be used to provide an approximate low-dimensional representation of CSHP (see below).

2.3.2 Meaning of axes

Considering the CSPH (which is a metric space) through a mapping method allows providing axes to the space. Ones can wonder the meaning of such axes. First, axes that come from non-linear mapping method have no specific meaning: any rotation or symmetry does not modify the map according to goodness criterion. Moreover, whatever the chosen axes, there is no a priori reason that could related axes to biological properties (as positions in the proteins for example). However, it cannot be exclude that a correlation between an axe and a biological property may be found by the reader.

2.4 Low-dimension representation of sequences as multidimensional points

As stated in section 2.3, sequences can be represented by points lying in a Euclidan multidimensional space. However, in sake of data exploration, an approximate visualization of such a space is desirable. In that purpose, we choose to use here a non-linear multidimensional scaling method (France and Caroll 2010, Lee and Verleysen 2007). Such representation cannot always preserve both geometry and topology of the original space, and distortions will appear (Lespinats and Aupetit 2011). In order to preserve local properties, the preservation of short distances is often favored. The information in maps is carried by distances between points. Here we used the data-driven high dimensional scaling (DD-HDS) criterion (Lespinats et al. 2007).

DD-HDS is preferred rather than

- linear methods such as principal component analysis (Pearson 1901) or classical multidimensional scaling (Torgerson 1968) which are linear. However, non-linear structures are not taken into account by linear methods meanwhile such types of structures are likely to appear in CSHP.

- Sammon's mapping (Gorelova et al. 2019; Sammon 1969) Indeed, even if Sammon's mapping is still widely used, it is a somewhat old method now known as prone to "false neighborhoods" (faraway items represented as neighbors).
- Non-metric methods such as SNE (Hinton and Rowies 2003) and tSNE (van der Maaten and Hinton, 2008). Currently, such methods are probably among the most popular methods in mapping community. These methods are efficient for preserving neighborhood relationships by considering a softmax transformation of distances. However, for that reason distances are often highly distorted. Here, for CSHP representation, distances are of main concern.

DD-HDS is designed to seek for a set of items in a low-dimensional output space that minimize

$$\varsigma = \sum_{i<j} |d_{ij} - d'_{ij}| . W_{ij} \quad (1)$$

$$\text{With } W_{ij} = 1 - F\left(\frac{\min\left(d_{ij}, d'_{ij}\right) - \left(mean_{i<j}(d_{ij}) - 2\times(1-\lambda)\times std_{i<j}(d_{ij})\right)}{2\times\lambda\times std_{i<j}(d_{ij})}\right) \quad (2)$$

Where, d_{ij} is the distance between sequences i and j; d'_{ij} is the Euclidean distance between associated items i and j in the "protein representation space" (*i.e.* output space); F is the cumulative distribution function for the Normal distribution N (0,1) and λ is a user-set parameter between 0 and 1. ς is the evaluation of the preservation of distances: difference between original and represented distances are compared (term $|d_{ij} - d'_{ij}|$) according to the weight W_{ij} which is high if items *i* and *j* are close in the original or in the representation space. Here we set $\lambda = 0.1$ as advised in (Lespinats et al. 2007). The optimization is performed by "force directed placement" (Morison et al 2003). Here again, axes have no specific meaning and maps must be considered as invariant by translation, rotation and symmetry (Degret and Lespinats 2018).

2.5 Simultaneous Representation of the Phylogenetic Relations in the Sequence Space

The phylogenetic trees (in Newick format, Lemey et al. 2009) obtained from the maximum likelihood method PhyML were depicted in the protein representation sequence space by linking the points accordingly. The internal nodes are located in the protein representation space iteratively. Each new parent node is located at the barycenter of its two children in the representation space, weighted by the parent-child distances in the graph. Links between points show the phylogenic tree in the protein representation space and thus are expected to represent the trajectory of the sequences along their evolutionary path.

(Stahnke et al. 2016) also represent trees directly onto data maps (not necessary biological data, but data from any field). However, in their work, trees code for data proximity and show clusters, conversely to our article where trees inform on distances between biological sequences and consequently are expected to show the phylogenetical relationships. Moreover, a phylogenetic tree-based color-code is also provided using ColorPhylo (Lespinats and Fertil 2011). A unique color is associated to each species according to its position in the phylogenetic tree. In that purpose, distances between species in the unrooted tree are mapped onto a 2D space which are related to the base of the HSV color-space (*i.e.* H and S components). Proximity between two species in terms of color informs on their proximity from phylogenetic point of view. Consequently, colors carry the same information as phylogenetic tree.

3. Results

3.1 para-aminobenzoic acid

Tetrahydrofolate (THF) and its derivatives, known as folates or B9 vitamins, are essential elements in the metabolism of all living organisms (Rébeillé et al. 2006). The THF molecule is

composed of a pterin moiety, a para-aminobenzoic acid (pABA) and a glutamate tail. While animals largely depend on their dietary sources for folate supply, bacteria, fungi and plants synthesize folates de novo. In plants, THF is assembled in mitochondria but its whole biosynthesis is localized in three subcellular compartments, the mitochondria, the plastids and the cytosol. All folate-producing organisms are synthesizing pABA in three steps: (i) the extraction of ammonia from glutamine (which is done by the PabA enzyme in *E. coli*), (ii) the condensation of this ammonia on chorismate to form 4-amino-4-deoxychorismate (which is done by the PabB enzyme in *E. coli*) and (iii) the removal of the pyruvate moiety to form pABA, a reaction catalyzed by the aminodeoxychorismate lyase (PabC in *E. coli*). In some species, the synthesis of 4-amino-4-deoxychorismate is catalyzed by a bifunctional enzyme regrouping the activities PabA and PabB. The first N-terminal domain is the glutamine amidotransferase (GAT) homologous to *E. coli* PabA whereas the second domain is the aminodeoxychorismate synthase (ADCS) homologous to *E. coli* PabB, (Basset et al. 2004; Camara et al, 2011). This is the case for fungi (Edman et al. 1993; James et al. 2002), land Plants (Camara et al. 2011) and some protozoans (Triglia and Cowman 1999). In higher plants, pABA is synthesized in plastids. A first analysis of the ADCS sequences homologous to the *Arabidopsis thaliana* shows that the domain architecture (*i.e.* mono- or bifunctional) can be very different across the various taxons but also within a given taxon. For example, the Eubacteria phylum exhibits both monofunctional (only the PabB domain) and bifunctional (presence of both GAT and ADCS domains) proteins. In most species, the ADCS domain can be subdivided in two sub-domains, the chorismate binding region and a region containing a sequence similar to the Anthranilate synthase I superfamily domain (table 1).

3.2 Phylogeny of chorismate binding region

The phylogeny based on the maximum likelihood for the chorismate binding region of the ADCSs domain from 32 species representative of the various kingdoms is depicted on figure 2.

As seen in this figure, almost all the monofunctional enzymes (B+C) are grouped in a same clade, except for the branch containing the two firmicutes *Bacillus subtilis* and *Clostridium difficile* which cluster within the bifunctional group (A+B+C). This figure also shows that most of the branches display a high branch support score except for the branch containing *Anabaena* and *Toxoplasma gondii*. In order to get a better idea of the relative distance between these various sequences, we represented the maximum likelihood phylogeny in a three-dimensional protein space which is a 3D projection of the original sequence space (figure3, 4 and 5). Figure 3 shows the decreasing of the cost function for DD-HDS (equation 1) from the initial map (obtained with classical MDS (Torgerson, 1965)) to the final map (logarithmic scale) and shows that the projected space represent proximity relationships between sequences in the CSHP much better than classical MDS.

From the figures 4 and 5 (with and without colorPhylo), it can be seen that monofunctional proteins occupy mainly two regions of the sequence space whereas the bifunctional enzymes are more widely spread out, suggesting that the bifunctional proteins have a higher degree of freedom (resp. a greater number of residues) that allows a deeper exploration of the space of possibilities (resp. the regions of the CSHP), maybe providing a higher flexibility and higher fitness to cope with their environment. Most of the species belonging to a same kingdom are more or less grouped together, as shown by the maximum likelihood phylogeny in figure 2. However, *Toxoplasma gondii*, an apicomplexan, was positioned far from other Apicomplexa (*Plasmodium*) but close to organisms such as those in the branch *Corynebacterium glutamicum*/*Teredinibacter turnerae*, indicating strong similarities between the sequences. These resemblances may explain why the maximum likelihood phylogenetic method failed to position *Toxoplasma gondii* in the tree with a good support value, and the same holds probably true for *Anabaena*. In other words, the trajectory of the Toxoplasma sequence along its evolutionary path could have led this sequence “near” the group *Corynebacterium*

glutamicum/*Teredinibacter turnerae* even if no common evolutionary/mutational history could be found between these two regions of the sequence space. This confirms the idea that two points relatively close in the sequence space are not necessarily close from an evolutionary point of view (Gorelova et al., 2019). One possible explanation can be that fitness value for the proteins under consideration can be very low in the space separating these sequences. Interestingly, it seems that all the monofunctional PabB are located in the same regions of the projected sequence space: top or down on the figures 4 and 5. This is also highlighted for the two Firmicutes *Clostridium difficile* and *Bacillus subtilis* despite the fact that they were associated with the *Plasmodium* group with good bootstrap values by the maximum likelihood phylogeny (figure 2). Although these two species adopted a completely different evolutionary path compared with the other prokaryotes, they kept some sequence similarities with the other members of their kingdom even if they are exploring new regions of the sequence space. From this point of view, it can be observed that most of the taxa, like Embryophyta, fungi or Archeabacteria, are localized in well-defined regions of the sequence space, but others, like Apicomplexa and Bacteria, spread onto larger parts of the sequence space. Following the ideas of Adami et al. (2000) and Povolotskaya and Kondrashov (2010), these organisms exhibit evolutionary paths that seem to cross the whole space to occupy new undiscovered regions. These 'pioneering' behaviors could possibly be related to higher mutational rates. Indeed, Adami et al. (2000) showed that changes in sequence length (leading to an expansion of the sequence space) provide new spaces into which the environmental information and can be recorded together with the sequence complexity associated with this information. Nevertheless, complexity (and hence information, Shannon 1969) can be achieved only if the sequence distribution in the sequence space is non-homogeneous (otherwise, all random sequences could be functional). As a consequence, high mutational rate could be a way to provide non-homogeneous distribution of functional sequence.

4. Conclusion

In this paper, we demonstrate that it is possible to investigate the evolutionary trajectories in the sequence space by representing phylogenetic trees onto a projected sequence space. These projections provide useful information, when phylogenetic branches fail to be supported for instance by bootstrap values and help therefore to resolve apparent incongruences. They are also indicative of the degree of freedom that a given protein sequence might have within a clade, as exemplified with the Apicomplexa group that is apparently associated with a larger "exploration capacity" in the space of possibilities. In some cases, reported here, higher mutation rates starting from an ancestral sequence, but also higher frequencies of protein fusions seem to allow more dynamic displacements of protein sequences in this spatial representation and could explain why some unrelated sequences can appear close in simple phylogenetic trees. This last assumption could be tested by looking at phylogenetic reconstructions together with representations in the sequence space of other multifunctional proteins.

In future works, the proposed method will be used to explore the sub-variety of CSHP where sequences lie. It is indeed necessary to fully ascertain that the observed distances between proteins are not too much distorted during the projection of the highly multidimensional sequence space into a 3D map. Distortions may occur when the CSHP is mapped and such distortions may lead to wrong inferences: (i) false neighbours in the projected space or (ii) tears of the original sequence space leading to false distant points in the projected space. For that reason, a careful study of distortions location and severity is needed. A recent method able to consider 3D data (conversely to previous techniques) will be used in that goal (Colange B, Vuillon L, Lespinats S and Dutykh D, personal communication). Lastly, we will explore the possibility to increase the information brought by phylogenetic trees by adding proximity and

remoteness between sequences onto the sub variety in the CSHP, for example through an ad-hoc color-code.

5. Acknowledgments

This work was supported by the French National Research Agency (ANR-10-LABEX-04 GRAL Labex, Grenoble Alliance for Integrated Structural Cell Biology; ANR-11-BTBR-0008 Océanomics; ANR-15-IDEX-02 GlycoAlps and "Origin Of Life" Cross Disciplinary Projects of the Univ. Grenoble-Alpes).

REFERENCES

Abascal F, Zardoya R, Telford MJ (2010) TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. Nucleic Acids Res 38: W7-13.

Adami C, Ofria C, Travis CC (2000) Evolution of biological complexity. PNAS 97(9): 4463–4468.

Anisimova M, Gascuel O (2006) Approximate Likelihood-Ratio Test for Branches: A Fast, Accurate, and Powerful Alternative. Systematic Biology 55(4): 539-552.

Altschul SF, Gish W, Miller W, Myers EW, Lipman D J (1990) Basic local alignment search tool. J Mol Biol 215: 403-410.

Aurrecoechea C, Brestelli J, Brunk BP, Dommer J, Fischer S, Gajria B, Gao X, Gingle A, Grant G, Harb OS, Heiges M, Innamorato F, Iodice J, Kissinger JC, Kraemer E, Li W, Miller JA, Nayak V, Pennington C, Pinney DF, Roos DS, Ross C, Stoeckert CJ Jr, Treatman C, Wang H

(2009) PlasmoDB: a functional genomic database for malaria parasites. Nucleic Acids Res 37: D539-543.

Basset GJC, Quinlivan EP, Ravanel S, Rébeillé F, Nichols BP, Shinozaki K, Seki M, Adams-Phillips LC, Giovannoni JJ, Gregory III JF, Hanson AD (2004) Folate synthesis in plants: the p-aminobenzoate branch is initiated by a bifunctional PabA-PabB protein that is targeted to plastids. Proceedings of the National Academy of Sciences of the United States of America 101: 1496-1501.

Bastien O, Ortet P, Roy S, Maréchal E (2005) A configuration space of homologous proteins conserving mutual information and allowing a phylogeny inference based on pair-wise Z-score probabilities. BMC bioinformatics 6(1): 49.

Berardini TZ, Reiser L, Li D, Mezheritsky Y, Muller R, Strait E, Huala E (2015) The Arabidopsis Information Resource: Making and mining the "gold standard" annotated reference plant genome. Genesis 53: 474–485.

Bornberg-Bauer E, Chan HS (1999) Modeling evolutionary landscapes: Mutational stability, topology, and superfunnels in sequence space. Proceedings of the National Academy of Sciences 96(19): 10689–10694.

Camara D, Richefeu-Contesto C, Gambonnet B, Dumas R, Rébeillé F (2011) The synthesis of pABA: Coupling between the glutamine amidotransferase and aminodeoxychorismate synthase domains of the bifunctional aminodeoxychorismate synthase from Arabidopsis thaliana. Archives of Biochemistry and Biophysics: 505(1): 83-90.

Dayhoff MO (1976) The origin and evolution of protein superfamilies. Fed Proc 35: 2132–2138.

Dayhoff MO, Barker WC, Hunt LT (1983) Establishing homologies in protein sequences. Methods Enzymol 91: 524–545.

Degret F, Lespinats S (2018) Circular background decreases misunderstanding of multidimensional scaling results for naive readers. In MATEC Web of Conferences (Vol. 189, p. 10002). EDP Sciences.

DePristo MA, Weinreich DM, Hartl DL (2005) Missense meanderings in sequence

space: a biophysical view of protein evolution. Nature Rev Genet 6: 678- 687.

Dryden DTF, Thomson AR, White JH (2008) How much of protein sequence space has been explored by life on Earth? J R Soc Interface 5: 953–956.

Edman JC, Goldstein AL, Erbe JG (1993) Para-aminobenzoate synthase gene of Saccharomyces cerevisiae encodes a bifunctional enzyme. Yeast 9: 669-675.

Facco E, d'Errico M, Rodriguez A, Laio A (2017) Estimating the intrinsic dimension of datasets by a minimal neighborhood information. Scientific Reports 7: 12140.

Felsenstein J (1981) Evolutionary trees from DNA sequences: A maximum likelihood approach. Journal of Molecular Evolution 17(6): 368–376.

France SL, Caroll JD (2010) Two-Way Multidimensional Scaling: A Review. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, PP (99): 1-18.

Gignoux C, Silvestre-Brac B (2002) Mécanique, de la formulation lagrangienne au chaos hamiltonien. EDP Sciences, Grenoble.

Gorelova V, Bastien O, de Clerck O, Lespinats S, Rébeillé F, Van Der Straeten D (2019) Evolution of folate biosynthesis and metabolism across algae and land plant lineages. Scientific Reports 9(1): 5731.

Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O (2010) New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. Systematic Biology 59(3):307-21.

Hinton GE, Roweis ST (2003) Stochastic neighbor embedding. In: Advances in neural information processing systems, pp 857–864.

Holm L, Sander C (1996) Mapping the protein universe. Science. 273: 595–603.

James TY, Boulianne RP, Bottoli AP, Granado JD, Aebi M, Kües U (2002) The pab1 gene of Coprinus cinereus encodes a bifunctional protein for para-aminobenzoic acid (PABA) synthesis: implications for the evolution of fused PABA synthases. Journal of basic microbiology 42: 91-103.

Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Molecular Biology and Evolution 30:772-780.

Kondrashov DA, Kondrashov FA (2015) Topological features of rugged fitness landscapes in sequence space. Trends in Genetics 31(1) 24-33.

Koonin EV, Wolf YI, Karev GP (2002) The structure of the protein universe and genome evolution. Nature 420: 218-223.

Kumar S, Stecher G, Tamura K (2016) MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. Molecular biology and evolution 33: 1870-1874.

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG (2007) Clustal W and Clustal X version 2.0. Bioinformatics 23: 2947-2948.

Lee JA, Verleysen M (2007) Nonlinear dimensionality reduction. Springer, New York, NY, USA.

Lemey P, Salemi M, Vandamme AM (2009) The Phylogenetic Handbook. Cambridge Press, Cambridge, 2nd Edition.

Lespinats S, Verleysen M, Giron A, Fertil B (2007) DD-HDS: A method for visualization and exploration of high-dimensional data. IEEE Transactions on Neural Networks 18 (5): 1265-1279.

Lespinats S, Fertil B (2011) ColorPhylo: a color code to accurately display taxonomic classifications. Evolutionary Bioinformatics, 7, EBO-S7565.

Lespinats S, Aupetit M (2011) CheckViz: Sanity Check and Topological Clues for Linear and Non-Linear Mappings. Computer Graphics Forum 30:113-121.

Lukasz P. Kozlowski LP (2017) Proteome-pI: proteome isoelectric point database. Nucleic Acids Res 45: D1112–D1116.

Marchler-Bauer A, Lu S, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C, Fong JH, Geer LY, Geer RC, Gonzales NR, Gwadz M, Hurwitz DI, Jackson J D, Ke Z, Lanczycki CJ, Lu F, Marchler GH, Mullokandov M, Omelchenko MV, Robertson CL, Song JS, Thanki N, Yamashita RA, Zhang D, Zhang N, Zheng C, Bryant SH (2011) CDD: a Conserved Domain Database for the functional annotation of proteins. Nucleic Acids Res 39(D):225-229.

Maynard Smith J (1970) Natural selection and the concept of a protein space. Nature 225:563–564.

Morrison A, Ross G, Chalmers M (2003) Fast Multidimensional Scaling through Sampling, Springs and Interpolation. Inf Visualization 2: 68–77.

Neath AA, Cavanaugh JE (2012) The Bayesian information criterion: background, derivation, and applications. Wiley Interdisciplinary Reviews: Computational Statistics 4(2): 199-203.

Nei M, Kumar S (2000) Molecular Evolution and Phylogenetics. Oxford University Press, New York.

Nordberg H, Cantor M, Dusheyko S, Hua S, Poliakov A, Shabalov I, Smirnova T, Grigoriev IV, Dubchak I (2014) The genome portal of the Department of Energy Joint Genome Institute: 2014 updates. Nucleic Acids Res 42(1): D26-31.

Pearson K (1901) On lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 2(11): 559-572.

Povolotskaya IS, Kondrashov FA (2010) Sequence space and the ongoing expansion of the protein universe. Nature 465 : 922-927.

Pruitt KD, Tatusova T, Maglott DR (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 33(Database Issue): D501–D504.

Rambaut A (2009) FigTree version 1.4.4 [computer program]. http://tree.bio.ed.ac.uk/software/figtree/

Rébeillé F, Ravanel S, Jabrin S, Douce R, Storozhenko S, Van Der Straeten D (2006) Folates in plants: biosynthesis, distribution, and Enhancement. Physiologia Plantarum 126: 330–342.

Shannon CE, Weaver W (1949) The Mathematical Theory of Communication. Univ of Illinois Press, Urbana, IL.

Romero PA, Arnold FH (2009) Exploring protein fitness landscapes by directed evolution. Nat Rev Mol Cell Biol 10(12): 866–876.

Sammon JW (1969) A nonlinear mapping for data structure analysis. IEEE Trans Comput C-18(5): 401–409.

Setubal JC, Meidanis J (1997) Introduction to computational molecular biology. PWS, Boston.

Stahnke J, Dörk M, Müller B, Thom A (2016) Probing Projections: Interaction Techniques for Interpreting Arrangements and Errors of Dimensionality Reductions. IEEE Transactions on Visualization and Computer Graphics 22(1): 629–638.

Starr TN, Thornton JW (2016) Epistasis in protein evolution. Protein Science 25(7) : 1204-1218.

Tokuriki N, Tawfik DS. (2009) Stability effects of mutations and protein evolvability. Curr Opin Struct Biol 19: 596-604.

Torgerson WS (1965) Multidimensional scaling of similarity. Psychometrika 30(4): 379-393.

van der Maaten L, Hinton G (2008) Visualizing data using t-SNE. J Mach Learn Res 9: 2579–2605.

Triglia T, Cowman AF (1999) Plasmodium falciparum: a homologue of p-aminobenzoic acid synthetase. Experimental parasitology 92: 154-158.

Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ (2009) Jalview Version 2-a multiple sequence alignment editor and analysis workbench. Bioinformatics 25: 1189-1191.

Wright S (1931) Evolution in mendelian populations. Genetics 16: 0097–0159.

Yau SST, Mao WG, Benson M, He RL (2015) Distinguishing Proteins From Arbitrary Amino Acid Sequences. Scientific Reports 5: 7972.

Young G, Householder AS (1938) Discussion of a set of points in terms of their mutual distances.Psychometrika 3: 19–22.

Table 1. List of ADCS sequences. Superfamilies and Chorimate Binding Domain (CBD) limits was identified using Conserved Domain Search Service v3.15 (Marchler-Bauer et al. 2011) with an expect value threshold equal to 0.01 and a composition-based statistics adjustment. The superfamily column shows the protein domain organization of the corresponding protein with **A**: GAT_1 superfamilly, **B**: Anth_synth_I_N superfamily, **C**: Chorismate_bind superfamily.

ID In the text	Species	Taxa	Kingdom Group	Sequence ID	Superfamily	CBD limits
ArabTha	Arabidopsis thaliana	Brassicaceae	Embryophyta	AT2G28880	**A** + **B** + **C**	1630-2709
PopuTri	Populus trichocarpa	Salicaceae	Embryophyta	POPTR_770528	**A** + **B** + **C**	1675-2754
SolaLyc	Solanum lycopersicum	Solanaceae	Embryophyta	350535750	**A** + **B** + **C**	1636-2646
PhysPat	Physcomitrella patens	Funariaceae	Embryophyta	PHYPADRAFT_190627	**A** + **B** + **C**	1351-2430
ChlaRei	Chlamydomonas reinhardti	Chlamydomonadaceae	Chlorophyta	CHLREDRAFT_123116	**A** + **B** + **C**	1057-2106
OstrLuc	Ostreococcus lucimarinus	Mamiellaceae	Chlorophyta	OSTLU_42151	**A** + **B** + **C**	1129-2253
Ustimay	Ustilago maydis	Ustilaginaceae	Fungi	UM03729.1	**A** + **B** + **C**	1957-2901
CrypNeo	Cryptococcus neoformans	Tremellaceae	Fungi	CNBA5540	**A** + + **C**	1510-2391
NeurCra	Neurospora crassa	Sordariaceae	Fungi	NCU01210	**A** + **B** + **C**	1543-2355
SaccCer	Saccharomyces cerevisiae	Saccharomycetaceae	Fungi	YNR033W	**A** + **B** + **C**	1495-2334
CandGla	Candida glabrata	Saccharomycetaceae	Fungi	CAGL0M10934g	**A** + **B** + **C**	1462-2292

CandAlb	Candida albicans	Saccharomycetaceae	Fungi	CaO19.1291 ABZ1	**A + B + C**	1606-2481
AspeFum	Aspergillus fumigatus	Trichocomaceae	Fungi	AFUA_6G04820	**A + B + C**	1576-2433
AspeOry	Aspergillus oryzae	Trichocomaceae	Fungi	AOR_1_106114	**A + B + C**	1561-2418
ToxoGon	Toxoplasma gondii	Conoidasida	Apicomplexa	TGME49_002920	**A + B + C**	2002-2913
PlasFal	Plasmodium falciparum	Aconoidasida	Apicomplexa	PFI1100w	**A + C**	1999-2865
PlasCha	Plasmodium chabaudi	Aconoidasida	Apicomplexa	PC001334.02.0	**A + B + C**	1864-2712
GranBet	Granulibacter bethesdensis	Acetobacteraceae	Eubacteria	GbCGDNIH1_0760	**A + B + C**	1327-2091
KineRad	Kineococcus radiotolerans	Kineosporiaceae	Eubacteria	Krad_3391	**A + B + C**	1087-1830
OligCar	Oligotropha carboxidovorans	Bradyrhizobiaceae	Eubacteria	OCA5_c27510	**A + B + C**	1225-1989
TereTur	Teredinibacter turnerae	Teredinidae	Eubacteria	TERTU_2979	**A + B + C**	979-2052
PseuFlu	Pseudomonas fluorescens	Pseudomonadaceae	Eubacteria	Pfl01_1759	**B + C**	271-1320
ClosDif	Clostridium difficile	Clostridiaceae	Eubacteria	CD1446	**B + C**	259-1326
CoryGlu	Corynebacterium glutamicum	Corynebacteriaceae	Eubacteria	NCgl0955 Cgl0997	**A + B + C**	835-1854
EschCol	Escherichia coli	Enterobacteriaceae	Eubacteria	EcolC_1821	**B + C**	316-1341
SalmTyp	Salmonella typhimurium	Enterobacteriaceae	Eubacteria	STM474_1846	**B + C**	319-1344
BaciSub	Bacillus subtilis	Bacillaceae	Eubacteria	BSn5_11940	**B + C**	298-1368
AnabSpe	Anabaena sp.	Nostocaceae	Cyanobacteria	alr3443	**A + B + C**	1003-2058

HalofSp	Haloferax sp.	Halobacteriaceae	Archaea	CQR50807	**B + C**	364-1464
NatroGr	Natronobacterium gregoryi	Halobacteriaceae	Archaea	AFZ72178	**B + C**	370-1461
NatroPh	Natronomonas pharaonis	Halobacteriaceae	Archaea	CAI48492	**B + C**	331-1425
NatriPe	Natrinema pellirubrum	Halobacteriaceae	Archaea	AGB33048	**B + C**	379-1497

Figure 1. Principle of the method, for more details about the actual used software in this study, see the material and method part. (a) Can be done by classical multiple alignment algorithm. (b) and (b') Computing of a distance matrix between biological sequences can be done using the sequences themselves, properties of their biological products or (b') considering their evolutionary relationships. (c) Molecular Phylogeny Inference can be done using Maximum Likehood Method, Bayesians Methods or (c') Distance methods. (d) Dimensional Scaling. (e) representation of the Newick tree generated by (c) or (c') into the projected Sequence Space.

Figure 2. Phylogenetic trees of the protein sequences of the Chorismate Binding domain constructed using maximum likelihood method PhyML (see text). The selected model using Bayesian information criterion was GTR+G+I with 4 categories, gamma shape parameter estimate=1.257 and the proportion of invariable site estimate = 0.102. Nodes values represents the aLRT branch support. The superfamily domain architectures of the proteins are given on the right of the leaf label (see material and method and table 1): **A**: GAT_1 superfamilly, **B**: Anth_synth_I_N superfamily, **C**: Chorismate_bind superfamily. The domain and group colors code correspond to: green, Chlorophyta and Embryophyta; Blue area, bifunctional ADCS; Beige area, monofunctional ADCS (PabB).

Figure 3. Decreasing of the cost function for DD-HDS (equation 1) from the initial map (reach with classical MDS (Torgerson, 1965)) to the final map (logarithmic scale).

Figure 4. Inference of the evolutionary trajectories of the Chorismate Binding Domain sequences in the sequence space. DD-HDS mapping method offers a configuration of points in this multidimensional space that is representative of the observed distances. Axes are arbitrary and the representation is invariant by rotation or lateral symmetry. In this figure, the distance between two data points on the figure tends to display the distances between the species biological sequences. Links between points show the maximum likelihood phylogenic tree presented in figure 2. Blue branches: bifunctional ADCS. Beige branches: monofunctional PabB. The yellow branches represent the branches with the nul aLRT branch support in figure 2. Green letters, Embryophyta and Chrorophyta. Abbreviations for genus names: Arab Tha, Arabidopsis thaliana; Popu Tri, Populus trichocarpa; Sola Lyc, Solanum lycopersicum; Phys Pat, Physcomitrella patens; Chla Rei, Chlamydomonas reinhardti; Ostr Luc, Ostreococcus lucimarinus; Usti may, Ustilago maydis; Cryp Neo, Cryptococcus neoformans; Neur Cra, Neurospora crassa; Sacc Cer, Saccharomyces cerevisiae; Cand Gla, Candida glabrata; Cand Alb, Candida albicans; Aspe Fum, Aspergillus fumigatus; Aspe Ory, Aspergillus oryzae; Toxo Gon, Toxoplasma gondii; Plas Fal, Plasmodium falciparum; Plas Cha, Plasmodium chabaudi; Gran Bet, Granulibacter bethesdensis; KineRad, Kineococcus radiotolerans; Olig Car, Oligotropha carboxidovorans; Tere Tur, Teredinibacter turnerae; Pseu Flu, Pseudomonas fluorescens; Clos Dif, Clostridium difficile; Cory Glu, Corynebacterium glutamicum; Esch Col, Escherichia coli; Salm Typ, Salmonella typhimurium; Baci Sub, Bacillus subtilis; Anab Spe, Anabaena sp., Halof Sp, Haloferax sp.; Natro Gr, Natronobacterium gregoryi; Natro Ph, Natronomonas pharaonic; Natri Pe, Natrinema pellirubrum.

Figure 5. Same as figure 4, using ColorPhylo, a unique color is associated to each species according to its position in the original phylogenetic tree. Proximity between two species in terms of color (hue, saturation and value) informs on their proximity from the phylogenetic point of view. Links between points show the maximum likelihood phylogenic tree presented in figure 2. Blue branches: bifunctional ADCS. Beige branches: monofunctional PabB. The yellow branches represent the branches with the aLRT branch support value in figure 2. Green letters, Embryophyta and Chrorophyta. Abbreviations for genus names: Arab Tha, Arabidopsis thaliana; Popu Tri, Populus trichocarpa; Sola Lyc, Solanum lycopersicum; Phys Pat, Physcomitrella patens; Chla Rei, Chlamydomonas reinhardti; Ostr Luc, Ostreococcus lucimarinus; Usti may, Ustilago maydis; Cryp Neo, Cryptococcus neoformans; Neur Cra, Neurospora crassa; Sacc Cer, Saccharomyces cerevisiae; Cand Gla, Candida glabrata; Cand Alb, Candida albicans; Aspe Fum, Aspergillus fumigatus; Aspe Ory, Aspergillus oryzae; Toxo Gon, Toxoplasma gondii; Plas Fal, Plasmodium falciparum; Plas Cha, Plasmodium chabaudi; Gran Bet, Granulibacter bethesdensis; KineRad, Kineococcus radiotolerans; Olig Car, Oligotropha carboxidovorans; Tere Tur, Teredinibacter turnerae; Pseu Flu, Pseudomonas fluorescens; Clos Dif, Clostridium difficile; Cory Glu, Corynebacterium glutamicum; Esch Col, Escherichia coli; Salm Typ, Salmonella typhimurium; Baci Sub, Bacillus subtilis; Anab Spe, Anabaena sp., Halof Sp, Haloferax sp.; Natro Gr, Natronobacterium gregoryi; Natro Ph, Natronomonas pharaonic; Natri Pe, Natrinema pellirubrum.