Algebraic Interpretations Towards Clustering Protein Homology Data

Fotis E. Psomopoulos, Pericles A. Mitkas

Algebraic interpretations towards clustering protein homology data

Fotis E. Psomopoulos[1*] and Pericles A. Mitkas[2]

[1] Center for Research and Technology Hellas
GR570 01, Thessaloniki, Greece

[2] Dept. of Electrical and Computer Engineering
Aristotle University of Thessaloniki
GR541 24, Thessaloniki, Greece

[*]corresponding author: fpsom@issel.ee.auth.gr

Abstract. The identification of meaningful groups of proteins has always been a principal goal in structural and functional genomics. A successful protein clustering can lead to significant insight, both in the evolutionary history of the respective molecules and in the identification of potential functions and interactions of novel sequences. In this work we propose a novel metric for distance evaluation, when applied to protein homology data. The metric is based on a matrix manipulation approach, defining the homology matrix as a form of block diagonal matrix. A first exploratory implementation of the overall process is shown to produce interesting results when using a well explored reference set of genomes. Near future steps include a thorough theoretical validation and comparison against similar approaches.

1 Introduction

In the era of Big Data, the quest for identifying hidden patterns and relationships is becoming an ever increasingly demanding objective, but at the same time, an imperative goal for most researchers. This situation holds particularly true in the fields of structural and functional genomics, where the need to assign potential functions and interactions to a rapidly expanding number of novel protein sequences is increasingly evident [1] [2]. There exist several algorithms in literature that address the issue of protein data clustering, ranging from generally applicable approaches [3] [4] [5], to highly specialized algorithms tailored for specific studies (i.e. studies focused on particular species [6], sets of genomes [7] or groups of molecules [8]).

A common concept in the vast majority of the clustering algorithms is "protein homology", i.e. the inherent degree of similarity that is assigned to a pair of protein sequences after application of a pair-wise comparison algorithm such as BLAST [9]. This similarity metric is consequently used to define new measures of distance, in

adfa, p. 1, 2011.

order to produce the necessary data partitioning, and therefore, insight into the intrinsic organization of the data involved.

A second key issue in any given clustering algorithm is the number of partitions created. As is often the case with big data, the actual number of "correct" and meaningful clusters is either unknown or hard to evaluate. Therefore there are two main approaches towards this issue; either approximate the number of clusters through external algorithms or parameters (such as in the case of k-means), or allow the clustering algorithm to construct an arbitrary number of clusters, based on its inner design (such as in the case of the popular MCL algorithm [3]).

In this work we propose a new algorithm, where the distance metric and an estimate of the clusters to be constructed is directly interpreted from the protein homology data. The rest of the paper is organized as follows: first we define the key concepts and techniques used within the context of the clustering process. The next section outlines the proposed algorithm and formally defines the metrics used. We conclude with the application of the novel metric on a well studied set of target genomes.

2 Problem Outline and Definitions

Attempting to formally define the protein clustering problem, we first need to provide the definitions of the concepts involved.

Given a set of $\boldsymbol{n}$ protein sequences, the homology matrix, a key concept in any protein clustering algorithm, can be defined as an $\boldsymbol{n} \times \boldsymbol{n}$ matrix $\boldsymbol{H}$ as follows:

$$H = \begin{bmatrix} h_{1,1} & h_{1,2} & \cdots & h_{1,n} \\ h_{2,1} & h_{2,2} & & h_{2,n} \\ \vdots & & \ddots & \vdots \\ h_{n,1} & h_{n,2} & \cdots & h_{n,n} \end{bmatrix} \qquad (1)$$

where $\boldsymbol{h_{i,j}}$ is the expect value of the pair-wise sequence comparison of protein sequence $\boldsymbol{i}$ and $\boldsymbol{j}$, using the BLAST algorithm. The expect-value, or e-value, is defined as the number of hits (correct alignments) expected to emerge by chance when searching a database of a certain size. At this point it is important to note that matrix $\boldsymbol{H}$ is square but not necessarily symmetrical, i.e. $h_{i,j} \neq h_{j,i}$.

A second key aspect of the homology matrix is that it is inherently sparse, i.e $\left|\{h_{i,j} \neq 0, \forall i,j\}\right| \ll \left|\{h_{i,j} = 0, \forall i,j\}\right|$. This qualitative characteristic is quantified through the sparsity metric of the matrix, defined as:

$$s = \frac{\left|\{h_{i,j} \neq 0, \forall i,j\}\right|}{\left|\{h_{i,j} = 0, \forall i,j\}\right|} \qquad (2)$$

A third aspect of a homology matrix emerges when protein sequences across $\boldsymbol{k}$ genomes (where $\boldsymbol{k} > 1$) are included in the data set. The sparsity pattern of the matrix $\boldsymbol{H}$ reveals a structure reminiscent of a block diagonal matrix. By definition, a block diagonal matrix is a square diagonal matrix in which the diagonal elements are square matrices of any size (possibly even 1×1), and the off-diagonal elements are zero matrices, i.e.:

$$A = \begin{bmatrix} A_1 & 0 & \cdots & 0 \\ 0 & A_2 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & A_l \end{bmatrix} \text{ where each } A_i \text{ is a square matrix} \qquad (3)$$

However, the formal definition of a block diagonal matrix differs with regard to the case of a homology matrix produced from sequences across $\boldsymbol{k}$ genomes. The difference lies in the fact that the off-diagonal elements are not zero, but exhibit significantly higher sparsity percentage $\boldsymbol{s}$ compared to the diagonal elements.

With the above definitions, the problem we are attempting to address can be formally defined as follows: given a homology matrix $\boldsymbol{H}$, define an appropriate distance metric $\boldsymbol{m}$ applied directly on $\boldsymbol{H}$, which will be consequently used within an agglomerative clustering process in order to produce a segmentation of $\boldsymbol{H}$ based on the $h_{i,j}$ values.

3 Methods

The algorithm comprises three distinct stages. The first pre-processing stage transforms the pair-wise comparison data into a full homology matrix. This is a necessary step as the standard output of the BLAST algorithm cannot be readily used in matrix manipulations, mainly due to the missing values and the scoring system employed. The second stage uses the constructed matrix in order to identify the target number of clusters. Finally, the proposed distance metric is applied through a standard agglomerative clustering process.

At this point it must be noted that the proposed method does not aim to replace the BLAST algorithm or to provide similar functionality. Instead, by directly utilizing the output of BLAST, we aim to construct a singular analysis method that evaluates this information in the form of clusters.

3.1 Pre-processing stage

The pair-wise alignment algorithm BLAST, employs a scoring system based on the expect value (e-value). Therefore, by definition, given two sequences seq_i and seq_j, the respective e-value would be:

$$e_{i,j} = \begin{cases} \text{missing value} & \text{, when } seq_i \text{ cannot be aligned with } seq_j \\ a > 0 & \text{, when there exist a valid alignment} \\ 0 & \text{, when the two sequences are identical} \end{cases} \quad (4)$$

In order to produce the homology matrix $\boldsymbol{H}$, we apply the following transformation:

$$h_{i,j} = \begin{cases} 0 & \text{, when } e_{i,j} = NaN \\ -\log_{10}(e_{i,j}) & \text{, when } e_{i,j} > 0 \\ \text{large constant c} & \text{, when } e_{i,j} = 0 \end{cases} \quad (5)$$

This linear transformation in essence changes only the range of values that appear in matrix $\boldsymbol{H}$ without affecting the attributes and characteristics of the data involved. With regard to the large constant, in our case we have set $\boldsymbol{c} = 1000$, but any sufficiently large number can be used

3.2 Cluster number estimation

As stated earlier, one of the key issues in data clustering is the definition of the number of clusters to be produced. In our case, this number is estimated directly from the characteristics of the homology matrix $\boldsymbol{H}$.

In linear algebra terms, the use of a block matrix corresponds to having a linear mapping thought of in terms of corresponding sets of basis vectors. This can be further viewed as having separate direct sum decompositions of both the domain and the range of the matrix. By completeness purposes, for any arbitrary matrices $A_{m\times n}$ and $B_{p\times q}$, the direct sum of A and B is denoted by A $\oplus$ B and defined as:

$$\mathrm{A} \oplus \mathrm{B} = \begin{bmatrix} a_{1,1} & \cdots & a_{1,n} & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ a_{m,1} & \cdots & a_{m,n} & 0 & \cdots & 0 \\ 0 & \cdots & 0 & b_{1,1} & \cdots & b_{1,q} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & b_{p,1} & \cdots & b_{p,q} \end{bmatrix} \quad (6)$$

Given the fact that the matrix $\boldsymbol{H}$ is square, we can also interpret the mappings as an endomorphism of an **n**-dimensional space V, i.e. a linear map such as f: V $\rightarrow$ V. In that regard, the block structure is of importance as it corresponds to having a single direct sum decomposition on V.

However, we must take under consideration the fact that the homology matrix $\boldsymbol{H}$ is an approximation of a block diagonal matrix. In order to evaluate the potential num-

ber of blocks existent within $\boldsymbol{H}$, we employ the λ eigenvalues of the matrix as the functional characteristic degree for the final segmentation.

3.3 Distance metric

The final stage in the clustering process requires the definition of an adequate distance metric. There are two key points that should be taken under consideration:

- Both dimensions of the homology matrix directly correspond to an ordered list of protein sequences. For the purposes of this work, the ordering of the protein sequences is based on the relative position of the respective genes on the genome chromosome.
- As can be also surmised from Equation (5), the distribution of values of matrix $\boldsymbol{H}$ is heavily biased on two ends, corresponding to the two cases of protein sequence alignment; no similarity (hence the sparsity of matrix) and complete identity.

Therefore, special care has be taken to include those attributes into the proposed metric, as seen in the equation below:

$$dist\left(h_{i_1,j_1}, h_{i_2,j_2}\right) = \begin{cases} \sqrt[3]{(h_{i_1,j_1} - h_{i_2,j_2})^2} & , \sqrt[2]{\Delta i^2 + \Delta j^2} \le c_1 \,\&\&\, \Delta h \le c_2 \\ (h_{i_1,j_1} - h_{i_2,j_2}) + \sqrt[2]{\Delta i^2 + \Delta j^2} & , \sqrt[2]{\Delta i^2 + \Delta j^2} > c_1 \,\&\&\, \Delta h \le c_2 \\ \sqrt[2]{\Delta i^2 + \Delta j^2 + \Delta h^2} & , \sqrt[2]{\Delta i^2 + \Delta j^2} > c_1 \,\&\&\, \Delta h > c_2 \\ \sqrt[2]{|\Delta h^2 + (\Delta i^2 - \Delta j^2)|} & , \sqrt[2]{\Delta i^2 + \Delta j^2} \le c_1 \,\&\&\, \Delta h > c_2 \end{cases} \quad (7)$$

where $c_1 = \frac{n}{|\lambda|}$ and c_2 an arbitrary constant satisfying $\overline{\left(\frac{\sum h_{i,j}}{|\{h_{i,j} \neq \{0,c\}, \forall i,j\}|}\right)} < c_2 < c$.

The metric clearly defines four distinct states in the homology matrix $\boldsymbol{H}$:

1. closely located genes within the same range of similarity. In this case, as the genes are expected to be linked at some level, the distance takes into account only the homology values but with a bias towards smaller distance.
2. closely located genes with significant difference in homology. In this case, the respective genes are expected to belong to different functional groups, and therefore the distance is biased towards larger values.
3. distant genes within the same range of similarity. This is a very interesting case, as it should contain genes across different species that exhibit a high level of similarity.
4. distant genes with significant difference in homology. This is the most distant case of gene similarity, therefore the maximum distance is assigned.

The function shown in Equation (7) is used in the clustering process in order to produce the distance matrix and, consequently, the required clusters.

4 Results

In order to evaluate the effectiveness of the proposed metric, we employed a well-studied group containing the following five genomes ([10], [11]):

1. *Mycoplasma genitalium, G-37* [12] (Bacteria; Firmicutes; Mollicutes; Mycoplasmatales) 479 genes, COGENT code: MGEN-G37-01.
2. *Ureaplasma urealyticum, serovar 3* [13] (Bacteria; Firmicutes; Mollicutes; Mycoplasmatales) 613 genes, COGENT code: UURE-SV3-01.
3. *Streptococcus pyogenes M1, SF370* [14] (Bacteria; Firmicutes; Bacilli; Lactobacillales) 1696 genes, COGENT code: SPYO-SF3-01.
4. *Buchnera aphidicola, SG* [15] (Bacteria; Proteobacteria; Gamma-proteobacteria; Enterobacteriales) 545 genes, COGENT code: BAPH-XSG-01.
5. *Nanoarchaeum equitans, Kin4-M* [16] (Archaea; Nanoarchaeota) 563 genes, COGENT code: NEQU-N4M-01.

The phylogenetic relationships of the species is represented by the dendrogram in **Fig. 1**.

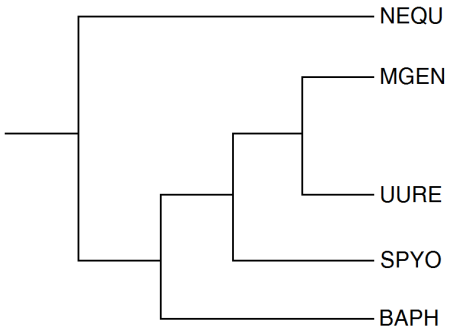


Fig. 1. Simplified dendrogram of the five species in the input dataset. The COGENT codes are used for genome representation.

In order to produce the initial matrix containing the expect value of the sequence similarities, we employed the BLAST algorithm with the default parameters. The homology matrix $\boldsymbol{H}$ is produced directly from the e-values, through the pre-processing step (Equation 5, where $\boldsymbol{c} = 1000$). A visual representation of the final matrix is shown in **Fig. 2** below.

It must be noted that **Fig. 2** showcases only the sparsity of the homology matrix, and does not take into account the actual values of the non zero elements. However, even this simplified representation is sufficient to evaluate the patterns that emerge from the genome-wide sequence comparison.

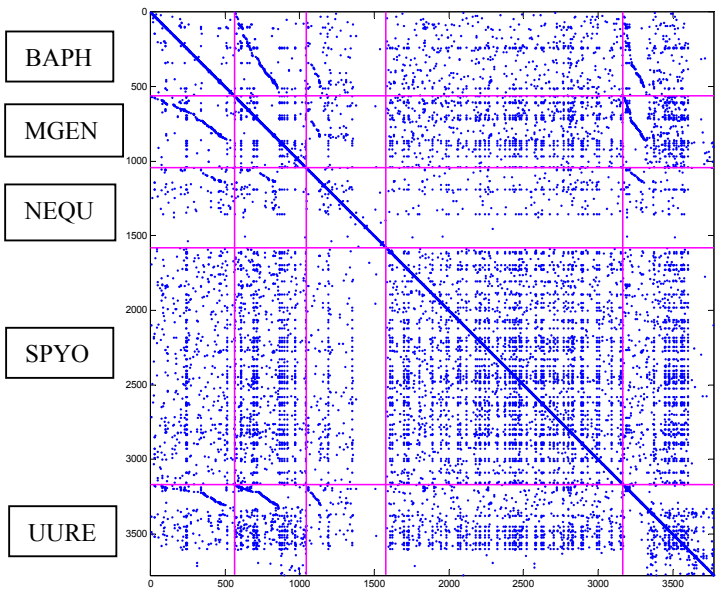


Fig. 2. Visual representation of the homology matrix. A non-zero value is denoted with a blue dot, whereas a zero value is denoted with a white dot. The purple lines show the limits of the five species in the dataset (same on both axes).

It is also critical to evaluate the initial assumption of the similarity between the homology matrix and a block diagonal matrix. To this end, **Fig. 3** shows the sparsity value of the sub-matrices, using Equation (2) for the metric evaluation.

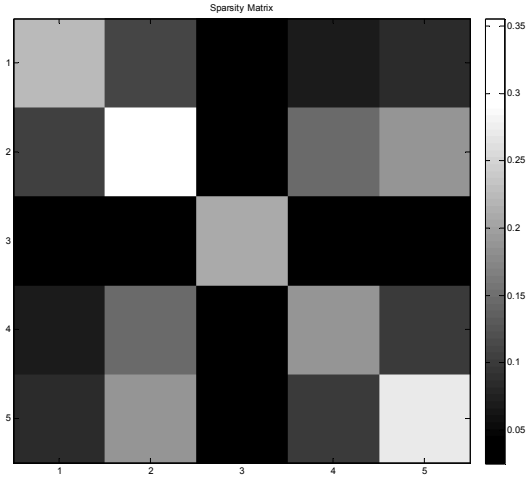
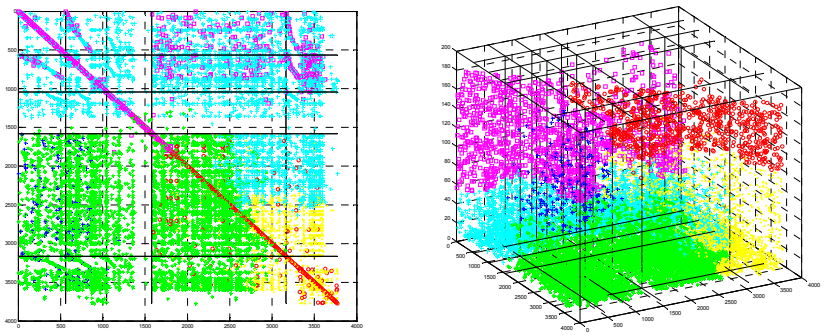


Fig. 3. Sparsity value of the homology matrix. Each sub-matrix corresponds to a pair of genomes.

Before applying the distance metric for the clustering process, we calculate the eigenvalues of the input matrix in order to set the number of clusters to be produced. Finally, inserting the proposed distance metric into the clustering process, we can produce a clustering of the protein sequences as shown in the following figures (**Fig. 4.** and **Fig. 5.**).



[a] [b]

Fig. 4. Clustering of the homology matrix across the three dimensions, where x and y correspond to the ordered protein sequences and z corresponds to the homology value of the respective sequence comparison. [a] shows the clustering on the two dimensions (i.e. ignoring the z axis) and [b] shows the clustering on all three dimensions.

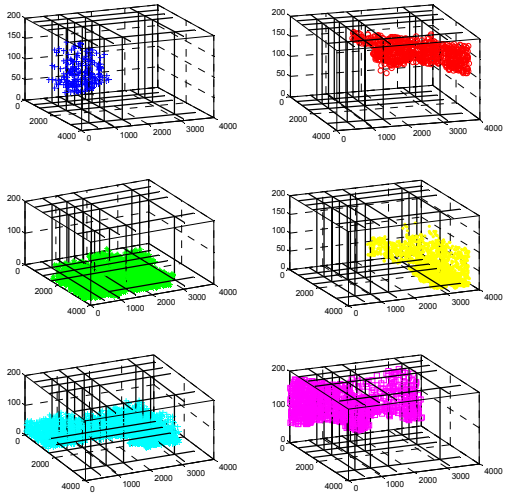


Fig. 5. Visualization of the six clusters the comprise **Fig. 4** employing the same color coding. It is evident that there is significant segmentation of the data points, both regarding the homology value and the different species the sequences belong to.

There are mainly two observations that can be made from the produced clusters; (i) the proposed distance metric allows for the discrimination of the different levels of homology, and (ii) the differentiation of the pair-wise genome comparisons (i.e. the different "cells" of the matrix) is not following closely the expected relationships.

Regarding the first observation, we can further infer from the results that there exist three distinct clusters containing the low value homology scores (green, yellow and cyan clusters), two correspond only to the high level similarity (red and magenta), and only one (blue cluster) that is constrained within a specific area of mid-level similarity. Specifically, this area corresponds to the two genome comparisons; *Streptococcus pyogenes* (SPYO) – *Mycoplasma genitalium* (MGEN) and *Streptococcus pyogenes* (SPYO) – *Buchnera aphidicola* (BAPH).

5 Discussion

In this work we propose an alternate distance metric for clustering protein sequences through a direct application on the corresponding homology matrix. Both the metric and the underlying concepts are based on the assumption that a homology matrix can be interpreted as a block diagonal matrix. This assumption is further explored and exploited through the estimate of the expected clusters and the definition of the distance metric.

We have provided some preliminary results by applying the proposed method in order to cluster data from a well-studied set of protein sequences. Although the results are very encouraging, by showing significant differentiation of the various levels

evident within the homology matrix, it is still a work in progress, requiring rigorous testing and validation. To this end, future work includes extensive comparison of the proposed method against similar algorithms, both from within the Machine Learning community and the Bioinformatics community.

Finally, an equally important issue to be explored is the scalability of the implemented algorithm. The problem of efficient and meaningful clustering of protein data is an open research issue which promises to become a key issue in next generation sequencing data analysis.

6 References

1. S. M. Williams and J. H. Moore, "Big Data analysis on autopilot?," *BioData Min.*, vol. 6, no. 1, p. 22, Jan. 2013.
2. R. Overbeek, M. Fonstein, M. D'Souza, G. D. Pusch, and N. Maltsev, "The use of gene clusters to infer functional coupling.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 96, no. 6, pp. 2896–901, Mar. 1999.
3. A. J. Enright, S. Van Dongen, and C. A. Ouzounis, "An efficient algorithm for large-scale detection of protein families.," *Nucleic Acids Res.*, vol. 30, no. 7, pp. 1575–84, Apr. 2002.
4. A. Sarkar, H. Soueidan, and M. Nikolski, "Identification of conserved gene clusters in multiple genomes based on synteny and homology.," *BMC Bioinformatics*, vol. 12 Suppl 9, no. Suppl 9, p. S18, Jan. 2011.
5. V. Miele, S. Penel, and L. Duret, "Ultra-fast sequence clustering from similarity networks with SiLiX.," *BMC Bioinformatics*, vol. 12, no. 1, p. 116, Jan. 2011.
6. R. Röttger, P. Kalaghatgi, P. Sun, S. D. C. Soares, V. Azevedo, T. Wittkop, and J. Baumbach, "Density parameter estimation for finding clusters of homologous proteins--tracing actinobacterial pathogenicity lifestyles.," *Bioinformatics*, vol. 29, no. 2, pp. 215–22, Jan. 2013.
7. D. E. Fouts, L. Brinkac, E. Beck, J. Inman, and G. Sutton, "PanOCT: automated clustering of orthologs using conserved gene neighborhood for pan-genomic analysis of bacterial strains and closely related species.," *Nucleic Acids Res.*, vol. 40, no. 22, p. e172, Dec. 2012.
8. J. Bonet, J. Planas-Iglesias, J. Garcia-Garcia, M. a Marín-López, N. Fernandez-Fuentes, and B. Oliva, "ArchDB 2014: structural classification of loops in proteins.," *Nucleic Acids Res.*, vol. 42, no. Database issue, pp. D315–9, Jan. 2014.
9. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool.," *J. Mol. Biol.*, vol. 215, no. 3, pp. 403–10, Oct. 1990.
10. S. Freilich, L. Goldovsky, A. Gottlieb, E. Blanc, S. Tsoka, and C. a Ouzounis, "Stratification of co-evolving genomic groups using ranked phylogenetic profiles.," *BMC Bioinformatics*, vol. 10, p. 355, Jan. 2009.
11. F. E. Psomopoulos, P. A. Mitkas, and C. A. Ouzounis, "Detection of genomic idiosyncrasies using fuzzy phylogenetic profiles.," *PLoS One*, vol. 8, no. 1, p. e52854, Jan. 2013.
12. Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA, et al. "The minimal gene complement of Mycoplasma genitalium". *Science* 270: 397–403. doi: 10.1126/science.270.5235.397, 1995.

13. Glass JI, Lefkowitz EJ, Glass JS, Heiner CR, Chen EY, et al. "The complete sequence of the mucosal pathogen Ureaplasma urealyticum". *Nature* 407: 757–762. doi: 10.1038/35037619, 2000.
14. Ferretti JJ, McShan WM, Ajdic D, Savic DJ, Savic G, et al. "Complete genome sequence of an M1 strain of Streptococcus pyogenes". *Proc Natl Acad Sci U S A* 98: 4658–4663. doi: 10.1073/pnas.071559398, 2001.
15. Shigenobu S, Watanabe H, Hattori M, Sakaki Y, Ishikawa H "Genome sequence of the endocellular bacterial symbiont of aphids Buchnera sp. APS". *Nature* 407: 81–86, 2000.
16. Waters E, Hohn MJ, Ahel I, Graham DE, Adams MD, et al. "The genome of Nanoarchaeum equitans: insights into early archaeal evolution and derived parasitism". *Proc Natl Acad Sci U S A* 100: 12984–12988. doi: 10.1073/pnas.1735403100, 2003.