ELECTIVE SUBJECT: ADVANCED TECHNIQUES IN MOLECULAR BIOLOGY
FINAL REPORT.

PREDICTION OF THE TERTIARY STRUCTURE OF A PROTEIN FROM A MODEL 3VMF PROTEIN.

Maria Alejandra Sierra G.
UNIVERSIDAD INDUSTRIAL DE SANTANDER
School of Biology

INTRODUCTION

Predicting the three-dimensional structure of a protein when only the amino acid sequence is known has been a major concern for many years[1]. Approaches to the prediction of the tertiary structure of a protein have been extended from purely *ab-initio* methods that rely exclusively on physical-chemical principles to homology methods based primarily on information available in the structure and sequence databases. The methods of folding prediction involve the identification of a structural model that more closely resembles the structure of a problem sequence. The objective of this laboratory work was to predict the tertiary structure of a problem protein *np_041900* from its amino acid sequence having as template the 3vmf protein. For this analysis, we used different software with different approaches such as 2D sequence comparison (HCA), Homology modeling (Swiss-Model), Protein folding recognition (I-TASSER) and prediction of *ab-initio* structure (I-TASSER). Also, using the model predicted by Swiss-model for the problem protein, an alignment was performed on the tertiary structure with the template protein in MATRAS-Markov Transition of Protein Structure Evolution-. In order to find conserved domains and to predict the possible function of the uncharacterized protein, we used the database PROSITE for protein families. Finally, the model was used to make a 3D visual alignment of the two proteins in Pymol. The primary and 3D sequences of the proteins were extracted from the NCBI and PDB databases respectively.

[1] Al-Lazikani et al, Protein structure prediction. Current Opinion in Chemical Biology 2001, 5:51–56

SEQUENCE OBTAINING

A sequence search was performed on NCBI to obtain the amino acid sequences of the 3vmf and np_041900 protein. Sequences were downloaded in FASTA format.

3vmf corresponds to an Archea elongation factor protein, *Aeropyrum pernix K1*, and np_041900 to a chloroplast elongation factor protein in *Euglena gracilis.*

SEQUENCE ALIGNMENT

The amino acid sequences of the two proteins were aligned in BLASTp in order to know their identity. A result of 33% was obtained.

COMPARISON OF 2D SEQUENCES WITH HCA

Cluster alignment was performed between hydrophilic and hydrophobic amino acids by hand and their identity was determined. An identity of 32.2% was calculated. Fig. 1.

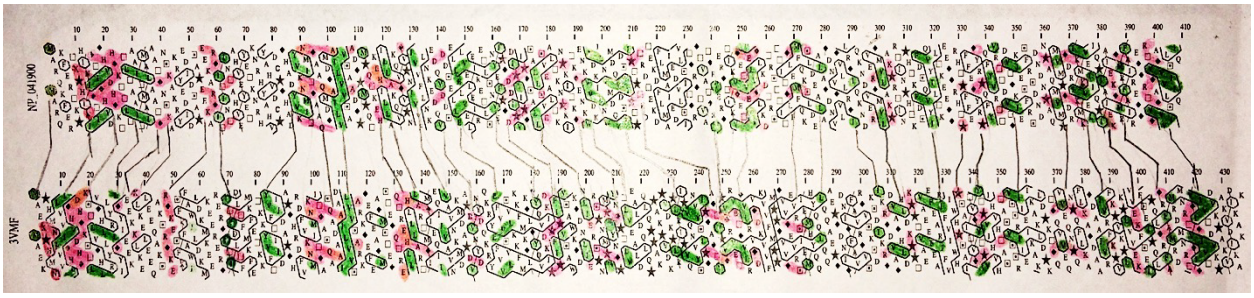


Fig 1. Alignment of HCA protein 3vmf and np_041900

OBTAINING MODELS PDB

In the PDB database a 3mvf protein search was performed and its three-dimensional structure was downloaded. Fig. 2.

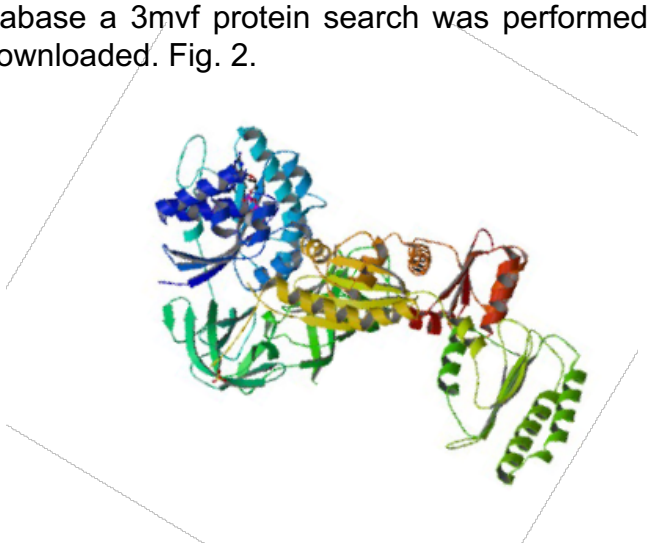


Fig 2. Three-dimensional structure 3mvf protein obtained from PDB

PREDICTION OF THE 3D MODEL IN SWISS-MODEL.

Since the tertiary structure of the np_041900 protein is unknown, a model in Swiss-Model had to be predicted from its amino acid sequence. Fig. 3.

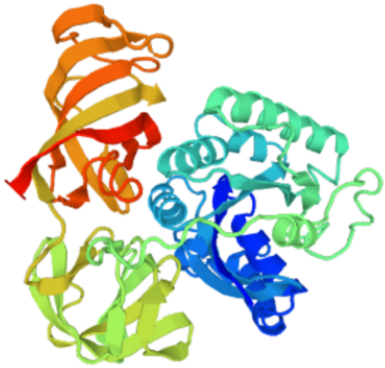


Fig 3. Model predicted by Swiss-Model for protein np_041900

PREDICTION OF 3D MODEL IN I-TASSER.

A prediction was made of the tertiary structure of the np_041900 protein in I-TASSER. Fig 4.

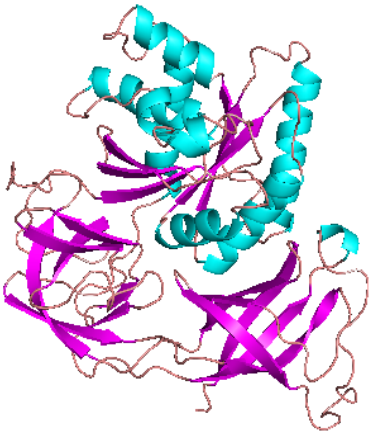


Fig 4. Image of the I-TASSER predicted tertiary protein model np_041900. Loops are represented in pink, β-Sheets in fuchsia and α-Helix in blue.

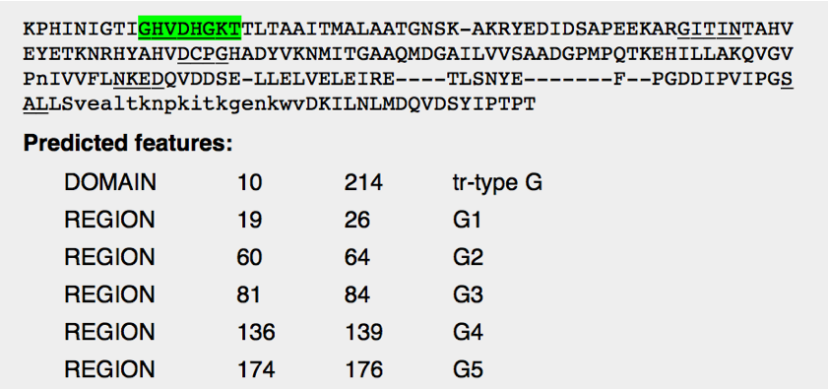
ALIGNMENT OF TERTIARY STRUCTURES

In MATRASS an alignment of the tertiary structures of the two proteins was made in order to know their identity. A result of 34.7% was obtained and a DRMS value of 1.82Å (mean square deviation -in angstrom- of the distances between the positions of the C^{beta} atoms of the aligned residues). The values between zero and three angstroms are accepted since they show a greater similarity between models [2].

[2] Chothial & M.Lesk. The relation between the divergence of sequence and structure in proteins. The EMBO Journal vol. 5. No. 4. Pp. 823, (1986)

SEARCH FOR CONSERVED DOMAINS

To predict the function of the uncharacterized protein, a possible identification was made of the protein families to which the sequence belongs. The PROSITE database was used and a conserved domain of type G_TR_1: *Translational (tr) -type guanine nucleotide-binding (G)* was identified. Fig. 5.


KPHINIGTIGHVDHGKTTLTAAITMALAATGNSK-AKRYEDIDSAPEEKARGITINTAHV
EYETKNRHYAHVDCPGHADYVKNMITGAAQMDGAILVVSAADGPMPQTKEHILLAKQVGV
PnIVVFLNKEDQVDDSE-LLELVELEIRE----TLSNYE-------F--PGDDIPVIPGS
ALLSvealtknpkitkgenkwvDKILNLMDQVDSYIPTPT

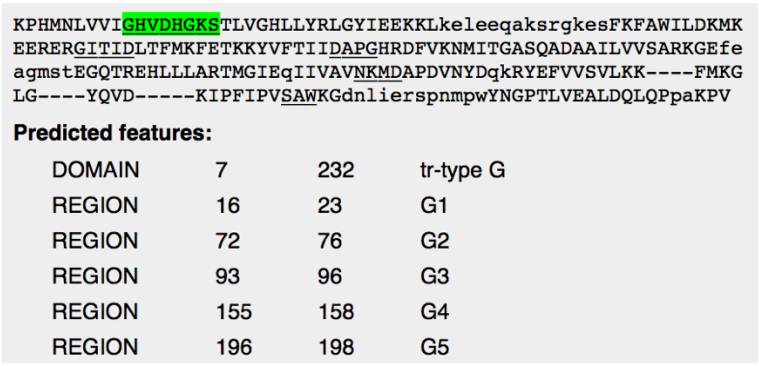
Predicted features:

DOMAIN	10	214	tr-type G
REGION	19	26	G1
REGION	60	64	G2
REGION	81	84	G3
REGION	136	139	G4
REGION	174	176	G5

Fig 5. Table of conserved regions and domains in protein np_041900. In green the conserved domain in the sequence is indicated

The loop of Guanosine Triphosphatases (GTPases) controls a multitude of biological processes, ranging from cell division, cell cycle and signal transduction to ribosome assembly and protein synthesis[3]. Translational GTPases (trGTPases) are a family of proteins in which GTPase activity is stimulated by the large ribosomal subunit. This family includes initiation of translation, elongation, and release factors.

In the same way, a domain search for the 3vmf template protein was carried out in order to verify the presence of this domain in both proteins. Fig. 6.


KPHMNLVVIGHVDHGKSTLVGHLLYRLGYIEEKKLkeleeqaksrgkesFKFAWILDKMK
EERERGITIDLTFMKFETKKYVFTIIDAPGHRDFVKNMITGASQADAAILVVSARKGEfe
agmstEGQTREHLLLARTMGIEqIIVAVNKMDAPDVNYDqkRYEFVVSVLKK----FMKG
LG----YQVD-----KIPFIPVSAWKGdnlierspnmpwYNGPTLVEALDQLQPpaKPV

Predicted features:

DOMAIN	7	232	tr-type G
REGION	16	23	G1
REGION	72	76	G2
REGION	93	96	G3
REGION	155	158	G4
REGION	196	198	G5

Fig 6. Table of conserved regions and domains in 3vmf protein. The conserved domain in the sequence is indicated in green.

[3] Leipe D.D., Wolf Y.I., Koonin E.V., Aravind L. Classification and evolution of P-loop GTPases and related ATPases. J. Mol. Biol. 317:41-72, (2002).

VISUALIZATION OF 3D STRUCTURES AND DOMAIN

The model of the np_041900 protein obtained from Swiss-Model was visualize in Pymol. The following tertiary structure was obtained in different views. Fig. 7

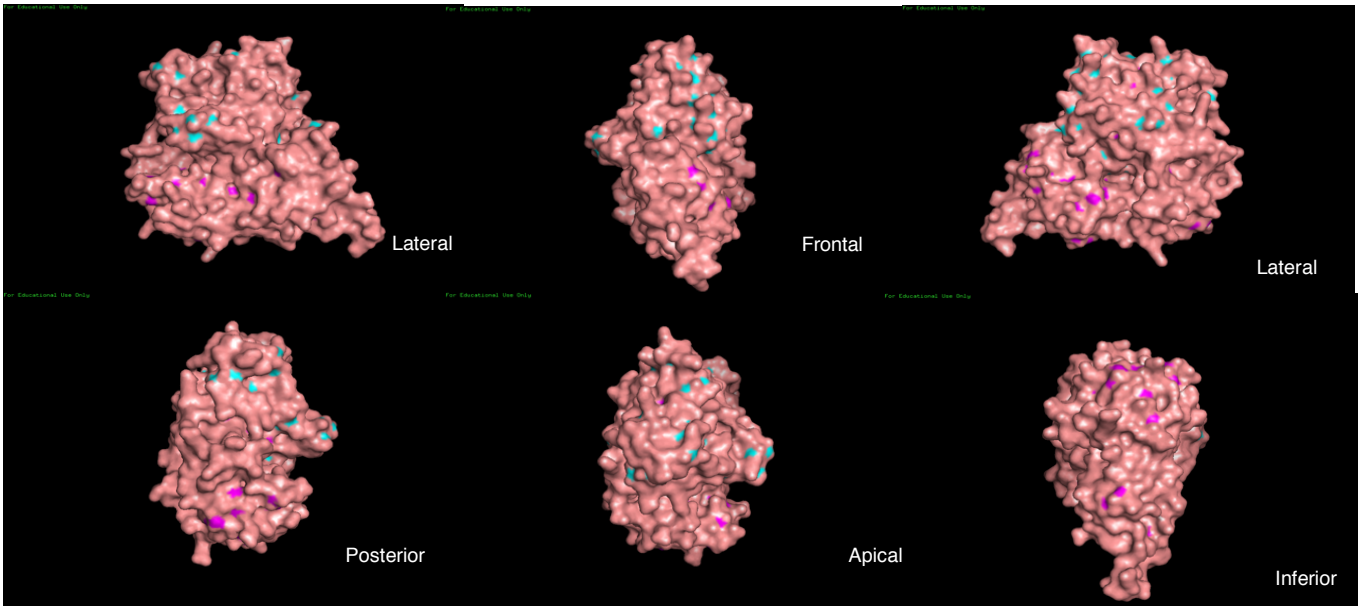


Fig 7. 3D image of the np_041900 protein model in different views. In pink the loops are represented, in fuchsia the β-sheets and in blue the α-helix.

Then, the conserved domain indicated by PROSITE was selected and the formation of a loop was observed as indicated in the literature [4].Fig. 8. In the same way, a domain for the 3vmf template protein was selected. Fig. 9.

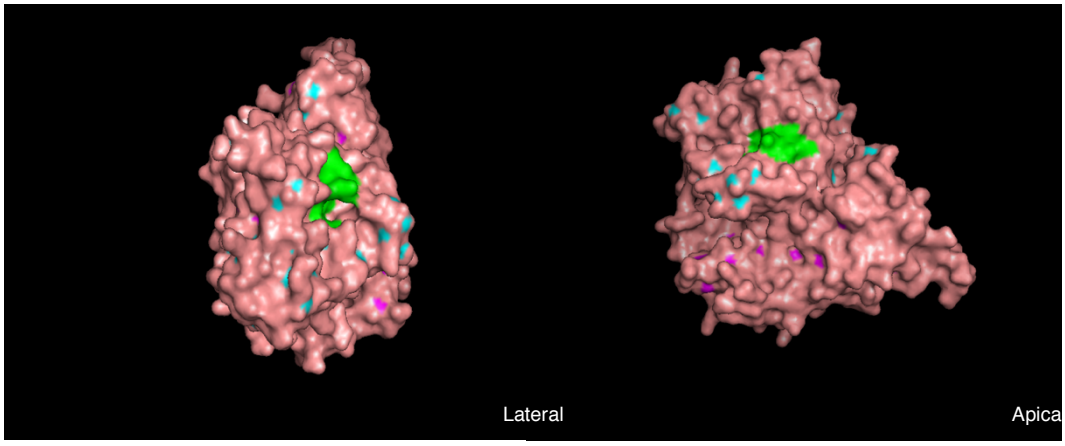


Fig 8. 3D image of the np_041900 protein where the conserved G_TR_1 domain is observed in green seen lateral and apically.

[4] Leibundgut M., Frick C., Thanbichler M., Boeck A., Ban N. Selenocysteine tRNA-specific elongation factor SelB is a structural chimaera of elongation and initiation facto. EMBO J. 24:11-22(2005).

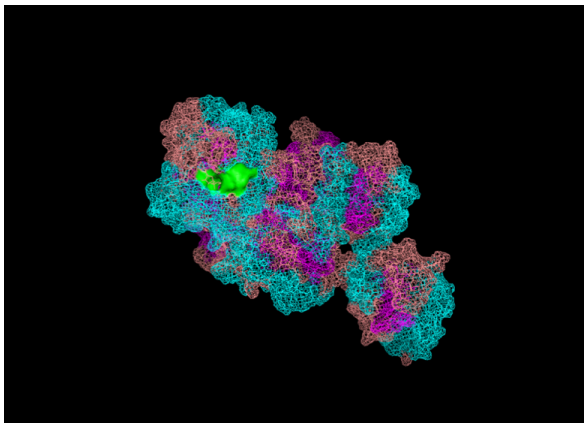


Fig 9. 3D image of the 3vmf protein where the G_TR_1 domain is observed in green.

ALIGNMENT OF TERTIARY PROTEIN STRUCTURES

The np_041900 model from Swiss-Model and the I-TASSER model was aligned with the 3vmf template protein obtained from PDB. Fig.10

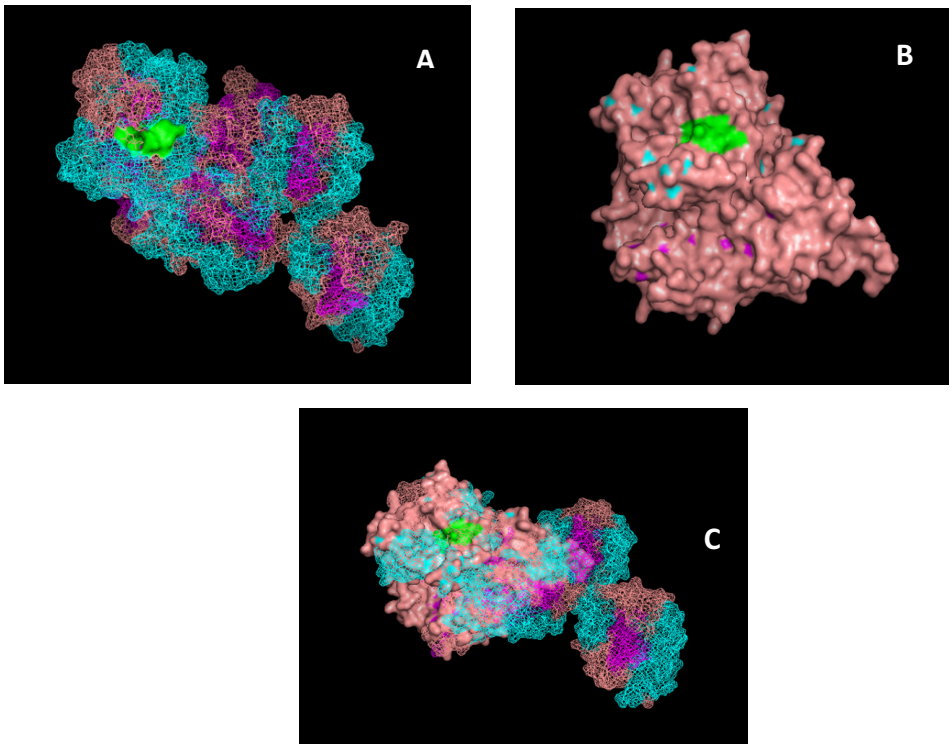


Fig 10. (A) 3D structure of the 3vmf template protein. **(B)** Model of protein problem predicted by Swiss-Model with its conserved domain identified in PROSITE. **(C)** 3D alignment of the two proteins.

Finally, an image of the tertiary structure of the two aligned proteins and the conserved domain was obtained. Fig. 11.

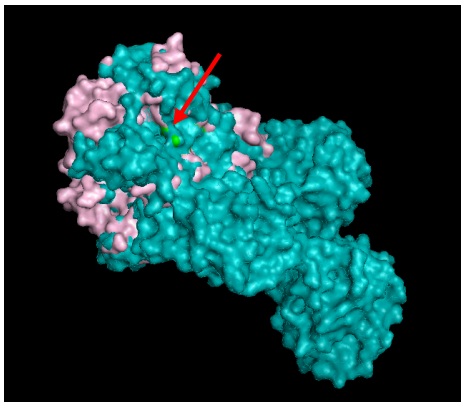


Fig 11. Representation of the tertiary structure of the two aligned proteins where the conserved (green) domain in the two proteins is indicated by a red arrow.

The RMS value as a result of the alignment of the Swiss-Model model and the 3vmf template protein was 1.145Å, while the alignment with the I-TASSER model indicates a value of 1.104Å. These values indicate a great similarity of the two structures[5].

ANALYSIS OF RESULTS.

Structural alignment is a type of sequence alignment based on the comparison of the shape that can predict the function of an unknown protein and also provide information about its evolutionary history[6]. The premises for predicting the structure of a protein based on a model are three:

1. Similar sequences adopt similar protein structures[7].
2. Many closely related sequences make up similar structures[8].
3. There are relatively few unique structures compared to the number of proteins in nature[9].

This is an important tool for comparing proteins with low similarity between their sequences, where evolutionary relationships between proteins cannot be easily detected by standard sequence alignment techniques.

The result of a structural alignment is an overlap of atomic coordinate sets, as well as a mean square distance (RMSD) between the basic structures of superimposed proteins. The RMSD of aligned structures indicates the divergences between them, the lower its value, the greater the similarity between the two proteins. In the case of alignments, RMSD values of 1,145Å were obtained with the Swiss-Model and 1.104Å with the I-

[5] O. Carugo. How root-mean-square distance (RMSD) values depend on the resolution of protein structures that are compared. J. Appl. Cryst (2003). 36, 125-128.
[6] Moult, J. A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. Current Opinion Structural Biology 15, (2005).
[7] Fischer, D., Rice, D., Bowie, J. U. Assigning aminoacid sequences to 3-dimensional protein folds. FASEB 10, 126-36, (1996).
[8] Chothia, C. and Lesk, A. M. The relation between the divergence of sequence and structure in proteins. EMBO J 5, 823-6, (1986).
[9] Wang, Z. X. A re-estimation for the total numbers of protein folds and superfamilies. Protein Eng 11, 621-6, (1998).

Tasser model in Pymol and 1,820Å with the Swiss-Model and 1.90Å with the I-Tasser model in MATRASS. These values indicate a great structural similarity of the models and 3vmf protein, which suggests that the np_041900 protein has a structure and function similar to that of 3vmf. 3vmf is a structural basis for the termination of the translation in a GTP complex in Archea[10].

One of the disadvantages of structural alignment is that, when using a predicted model to compare the problem protein, in the case of np_041900 with the known 3vmf protein, one cannot be sure how much this model differs from the actual structure of the protein. The literature recommends comparing the model predicted by different algorithms with the actual structure of the 3D protein, but in cases such as the one in which its tertiary structure was not obtained by crystallography, the structural alignment based on homology comparisons is a good prediction of function, structure, and evolutionary history[11].

CONCLUSIONS

A structural alignment is a type of sequences alignment based on the comparison of the protein structure. These alignments attempt to establish equivalences between two or more structures based on their three-dimensional conformation. It is also possible to perform a structural alignment with models produced by prediction methods such as I-Tasser and Swiss-Model. Although, the predicted models often require a structural alignment between the model and the actual known structure to evaluate the quality of the model.

Through the structural alignment of the models and the model protein, it can be said that these models resemble the real tertiary structure of the problem protein, and the structural similarity of an unknown protein and a model previously studied can assume the function that this protein.

BIBLIOGRAPHY

Adam Godzik. The structural alignment between two proteins: Is there a unique answer?. Protein Science Volume 5, Issue 7, pag 1325–1338. (1996)

Al-Lazikani et al, Protein structure prediction. Current Opinion in Chemical Biology 2001, 5:51–56

Chothial & M.Lesk. The relation between the divergence of sequence and structure in proteins. The EMBO Journal vol. 5. No. 4.

Fischer, D., Rice, D., Bowie, J. U. Assigning aminoacid sequences to 3-dimensional protein folds. FASEB 10, 126-36, (1996).

Kobayashi, K. et al. Structural basis for translation termination by archaeal RF1 and GTP-bound EF1alpha complex. Nucleic Acids Res. 40. (2012)

Leibundgut M., Frick C., Thanbichler M., Boeck A., Ban N. Selenocysteine tRNA-specific elongation factor SelB is a structural chimaera of elongation and initiation facto. EMBO J. 24:11-22(2005).

Leipe D.D., Wolf Y.I., Koonin E.V., Aravind L. Classification and evolution of P-loop GTPases and related ATPases. J. Mol. Biol.

Moult, J. A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. Current Opinion Structural Biology 15, (2005).

O. Carugo. How root-mean-square distance (RMSD) values depend on the resolution of protein structures that are compared. J. Appl. Cryst (2003). 36, 125-128.

Pp. 823, (1986)

Wang, Z. X. A re-estimation for the total numbers of protein folds and superfamilies. Protein Eng 11, 621-6, (1998).

[10] Kobayashi, K. et al. Structural basis for translation termination by archaeal RF1 and GTP-bound EF1alpha complex. Nucleic Acids Res. 40. (2012)

[11] Adam Godzik. The structural alignment between two proteins: Is there a unique answer?. Protein Science Volume 5, Issue 7, pag 1325–1338. (1996)