Lecture 03: Proteins Modelling

Dr. Alexis Salas Burgos
Marzo 2022

Topics

1.0 Protein Folding Indagation

1.1 Protein Folding and Misfolding

1.2 Levinthal Paradox

1.3 Energy Landscape

1.4 Free structure Prediction

1.5 AlphaFold

1.6 Comparative Modelling

1.7 Modeller

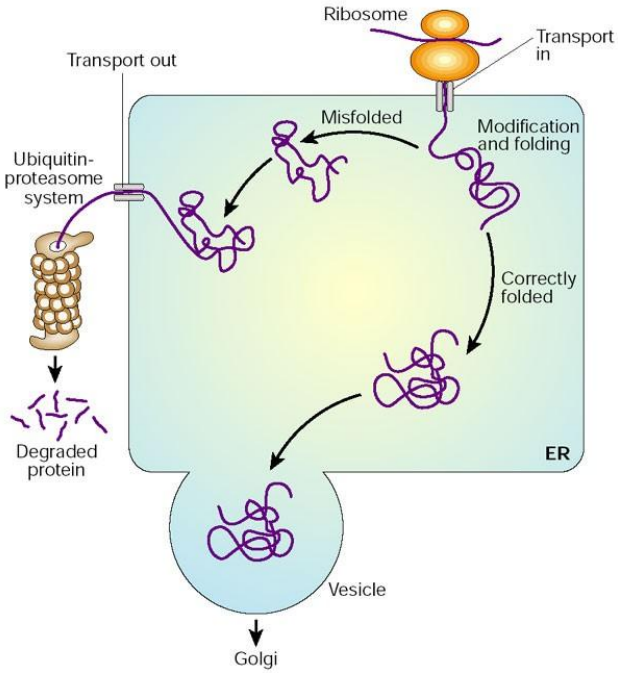
Overview

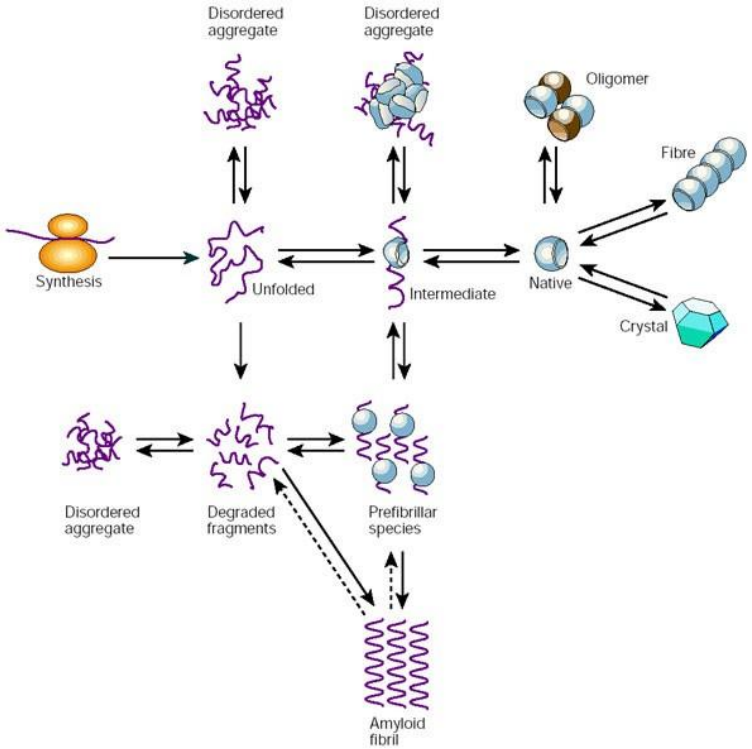
Para comprender las metodologías utilizados para la construcción de modelos tridimensionales de proteínas y aduiqir una capcidad análitica para su diseño y aplicación se deben comprender dos grandes principios:

- ¿Cómo las proteínas/moléculas se pliegan a sus estructuras 3D más estables ?
- ¿ Cómo la evolución de las secuencias influye en la relación estructura-función ?

Proteins Folding

1.1 Protein Folding and Misfolding





Dobson, Christopher M. 2003. “Protein Folding and Misfolding.” *Nature* 426 (6968): 884–90. https://doi.org/10.1038/nature02261.

1.1 Protein Folding and Misfolding

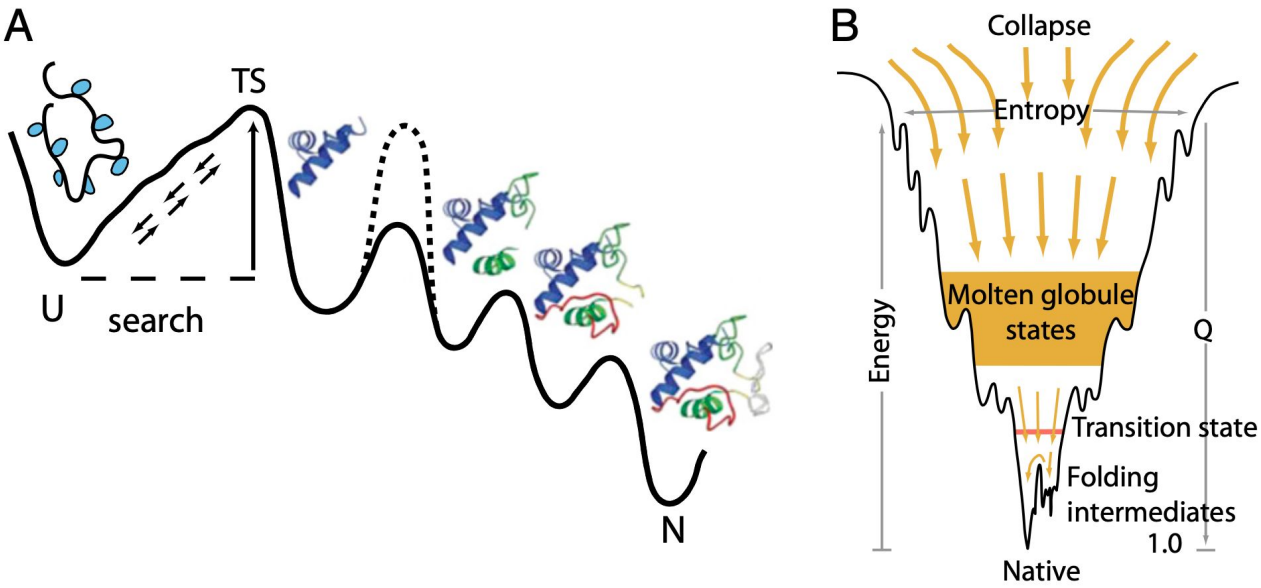


Fig. 1. (*A*) The classical view of a defined folding pathway, and (*B*) the new view of multiple routes through a funneled landscape. Reprinted with permission from ref. 13. Dashed line in *A* illustrates the insertion of an optional error-dependent kinetic barrier, which can affect some population fraction and not others and thus mimic multipathway folding.

Englander, S. Walter, and Leland Mayne. 2014. "The Nature of Protein Folding Pathways." *Proceedings of the National Academy of Sciences of the United States of America* 111 (45): 15873–80. https://doi.org/10.1073/pnas.1411798111.

1.1. Protein Fold Codon Effect

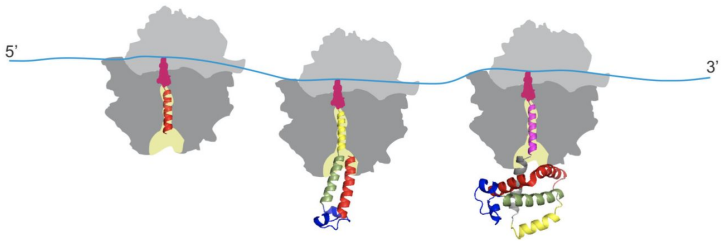


Fig. 1.
Co-translational folding is a stepwise vectorial process proceeding in most of the cases from the N- to C-terminal end of the growing polypeptide chain. The appearance of distinct folding intermediates is linked to the overall and local rates of translation. The process begins very early during polypeptide chain synthesis on the ribosome, with some secondary structure elements (such as e.g., alpha-helices) forming inside the ribosomal tunnel (and some tertiary structures forming inside the vestibule (lower/wider) region of the ribosomal exit tunnel. The tertiary structure of the protein is nearly formed at the end of the protein synthesis.

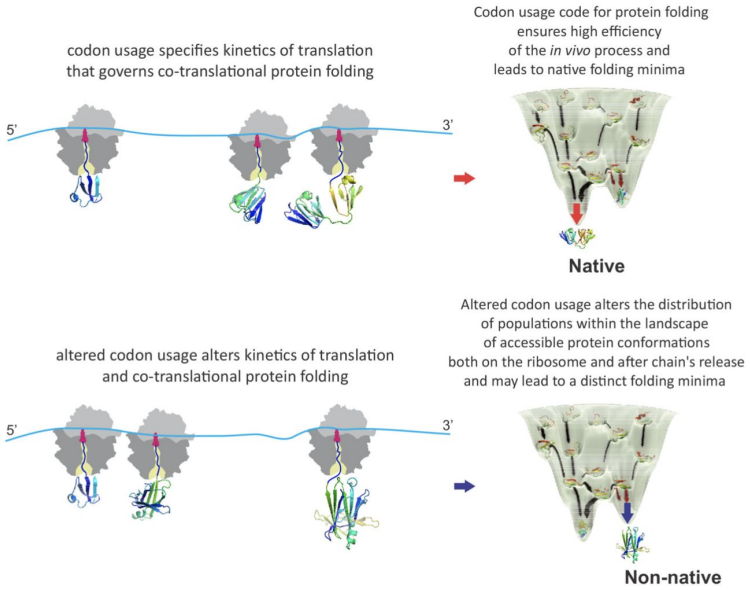


Fig. 4.
Genetic code governs protein folding in the cell. Appearance of the distinct folding intermediates is linked to the overall and local rate of translation specified by synonymous codon usage. Altered codon usage results in modified translation kinetics and folding events on the ribosome and after polypeptide chain release and may lead to altered final protein conformation and distinct folding minima.

Komar, Anton A. 2021. “A Code Within a Code: How Codons Fine-Tune Protein Folding in the Cell.” *Biochemistry. Biokhimiia* 86 (8): 976–91. https://doi.org/10.1134/S0006297921080083.

1.1 Protein Folding and Misfolding

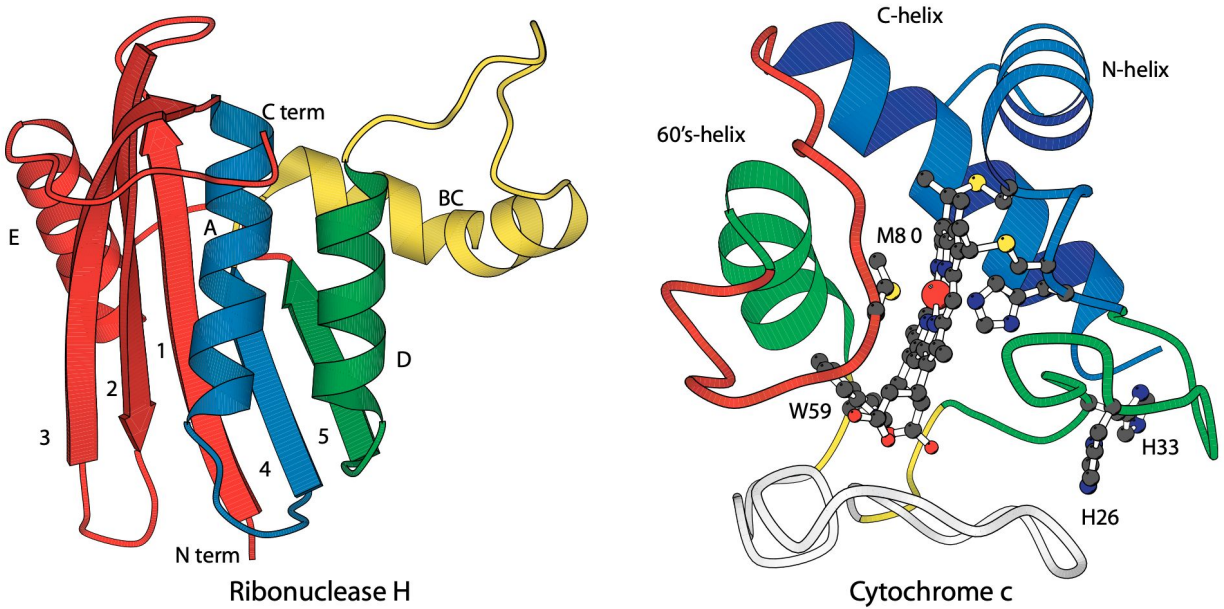
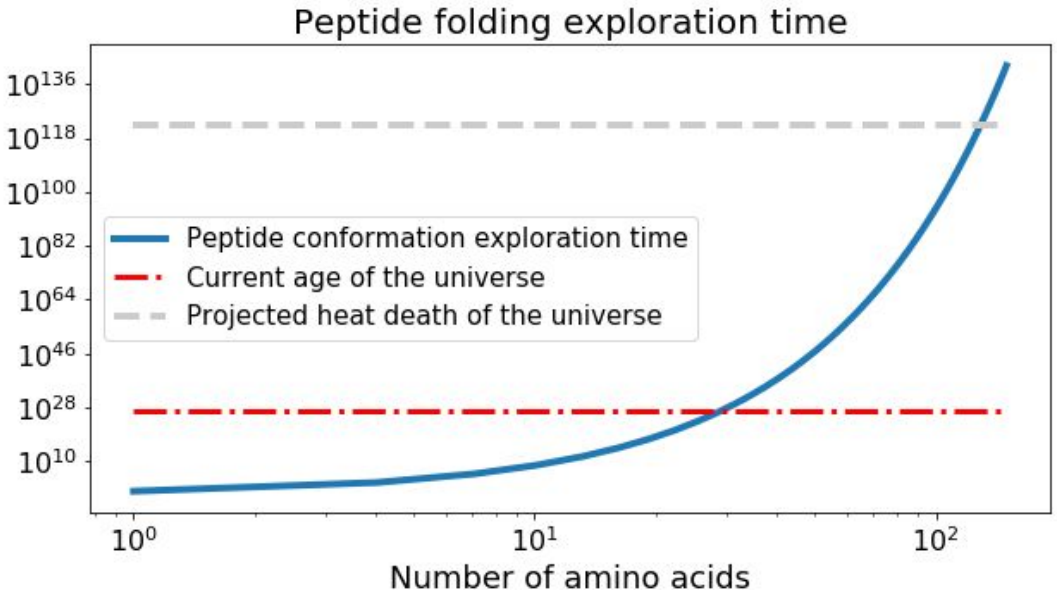


Fig. 6. The foldon construction of Ribonuclease H and Cyt c. The order of folding is blue, green, yellow, and red, and finally gray for the large bottom Cyt c loop.

Englander, S. Walter, and Leland Mayne. 2014. "The Nature of Protein Folding Pathways." *Proceedings of the National Academy of Sciences of the United States of America* 111 (45): 15873–80. https://doi.org/10.1073/pnas.1411798111.

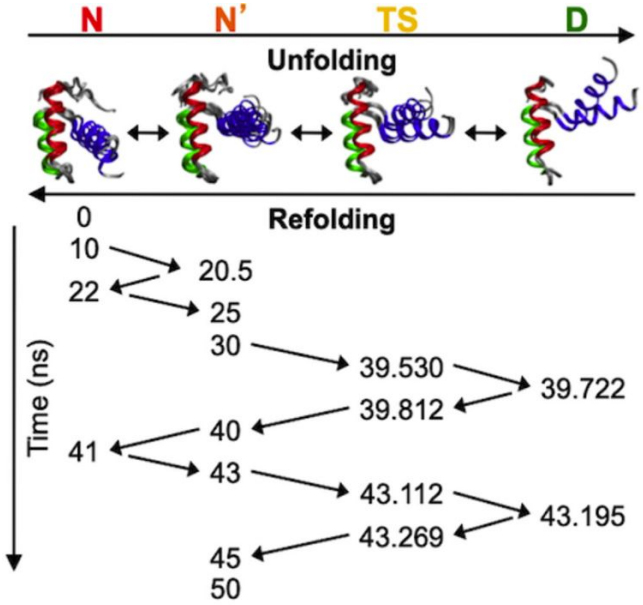
Levinthal's Paradox

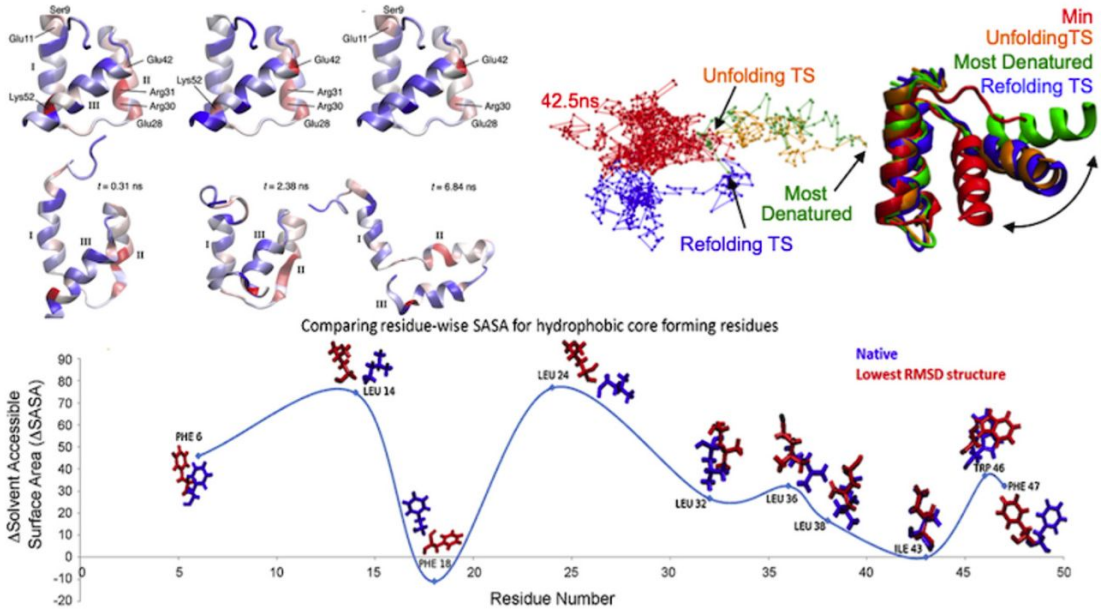


Levinthal (1969) Levinthal's Paradox. In Mossbauer spectroscopy in biological systems: proceedings of a meeting held at allerton house, DeBrunner J T P ME (ed), pp 22–24. Monticello, IL: University of Illinois
Google Scholar

"A Summary of DeepMind's Protein Folding Upset at CASP13 | Exxact Blog." 2019. *Exxact* (blog). June 11, 2019. https://blog.exxactcorp.com/deepminds-protein-folding-upset/.

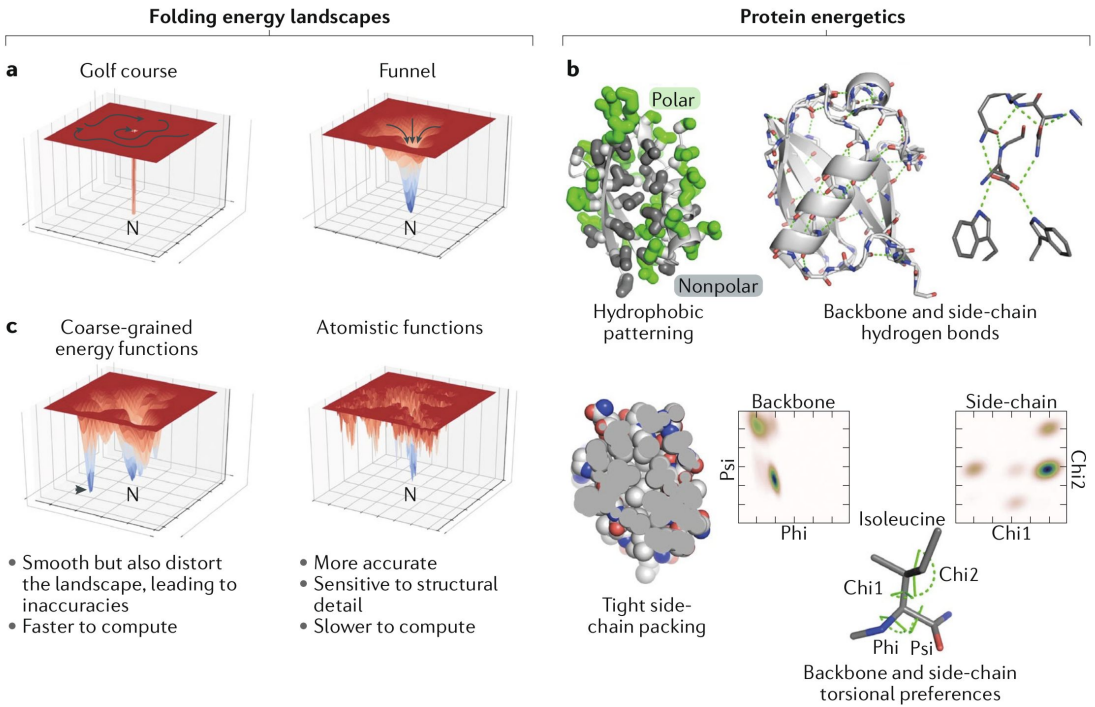
Folding and unfolding





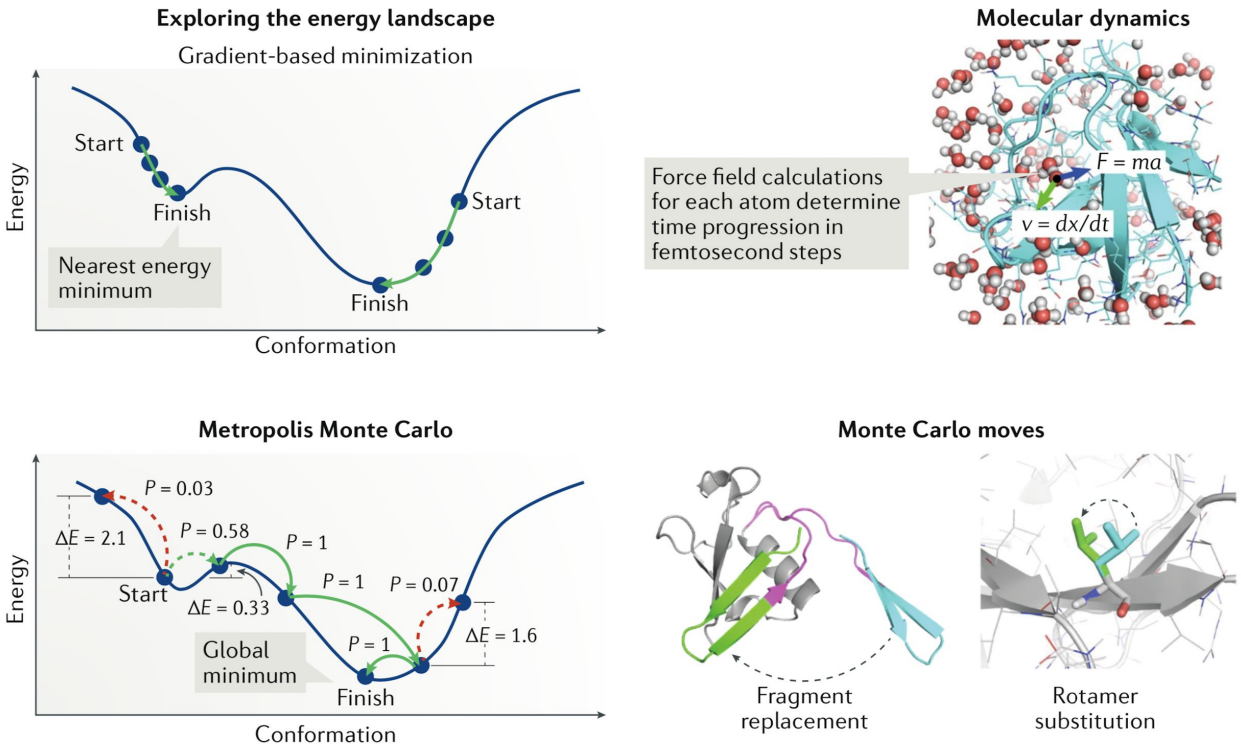
Ferina, Jennifer, and Valerie Daggett. 2019. “Visualizing Protein Folding and Unfolding.” *Journal of Molecular Biology* 431 (8): 1540–64. https://doi.org/10.1016/j.jmb.2019.02.026.

Protein Energy Landscape



Kuhlman, Brian, and Philip Bradley. 2019. “Advances in Protein Structure Prediction and Design.” *Nature Reviews Molecular Cell Biology*, August, 1–17. https://doi.org/10.1038/s41580-019-0163-x.

Protein Energy Landscape



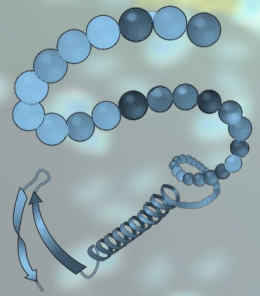
Kuhlman, Brian, and Philip Bradley. 2019. “Advances in Protein Structure Prediction and Design.” *Nature Reviews Molecular Cell Biology*, August, 1–17. https://doi.org/10.1038/s41580-019-0163-x.

Co-evolución

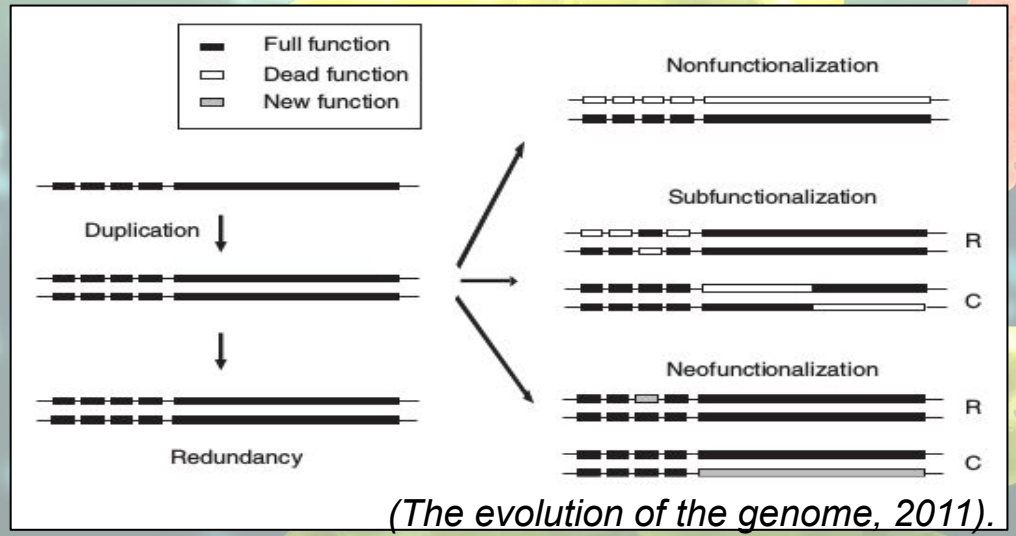
INTRODUCCIÓN

Proteínas

- Plegamiento
- Función
- Adaptación



«La conformación nativa de una proteína en su medio fisiológico está determinada por la totalidad de las interacciones interatómicas y, por tanto, por su secuencia aminoacídica» *(Anfinsen, C. B. 1973)*



(The evolution of the genome, 2011).

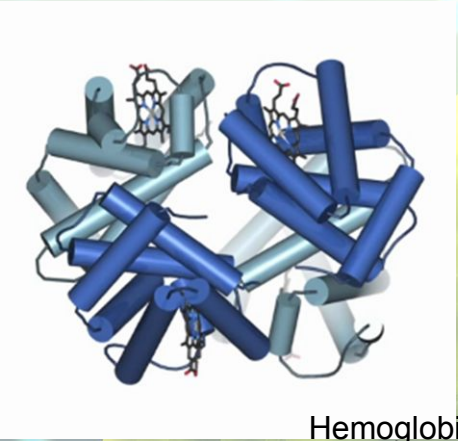
EL PROBLEMA DE LA FUNCIÓN

Richard Feynman, 1961.
The relation of Physics to other sciences

"One of the great triumphs of recent times within the last year…was at last to discover the exact arrangement in space…and the order…of some 56 or 60 amino-acids in a row…over a 1000 atoms…have been all located in a complex pattern for one, well now two, proteins…the first was hemoglobin."

"One of the sad parts of this is that for all we can tell, yet, we can't see anything from this pattern. We don't understand why it works the way it does because of that pattern…but of course that is the next problem."

Interacción Proteína-Ligando | Especificidad | Afinidad | Regulación



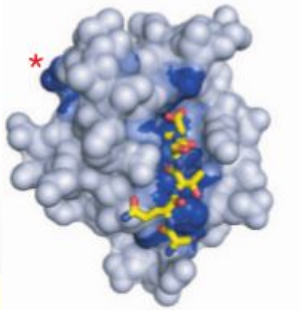
Hemoglobina. *(Preutz, 1959)*

¿Qué información de la secuencia es necesaria y suficiente para determinar las propiedades funcionales de una proteína?

LO QUE SABEMOS

Interacciones cooperativas ente residuos

- Dependientes de la geometría y distancia
- Unión de ligandos: Principio de proximidad espacial
- Interacciones funcionales: **locales** y **no locales**
- Patrones no son infebriles directamente desde estructura 3D



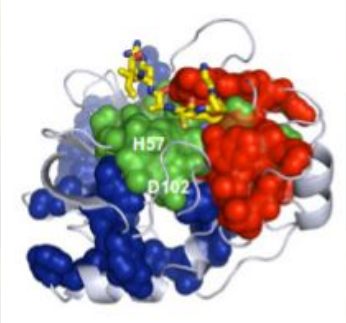
Dominio PDZ. *(McLaughlin, 2012).*

Caracterización de interacciones funcionales

Interacciones determinantes: Sólo una pequeña fracción de los aminoácidos de una proteína son determinantes para su función *(Thorn, K. S. , 1998)*

Modularidad: Grupos diferentes de residuos se asocian a funciones/propiedades diferentes en una proteína.

Conservación y co-evolución: Mantención de la estabilidad, plegamiento y función de la proteína.

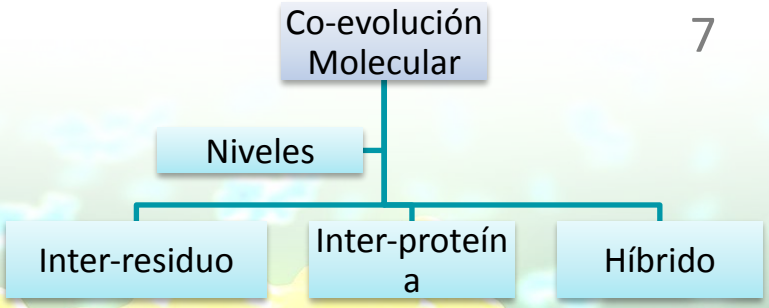


Sectores funcionales en tripsina de rata, PDB 3TGI *(Halabi, 2009).*

CO-EVOLUCIÓN MOLECULAR

Co-evolución: Interdependecia entre los cambios evolutivos de dos o más **entidades biológicas**.

Nivel molecular: Cambios coordinados en las secuencias de los **genes** y sus respectivos **productos proteicos.**



(De Juan, 2013).

Inter-especies → Molecular

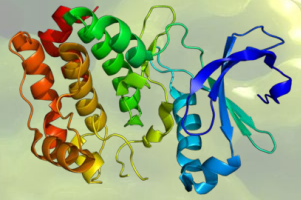
Darwin,1862

Dobzhansky, 1950

Ehrlich, 1964

*«Cambios evolutivos recíprocos entre especies **interactuantes**»*

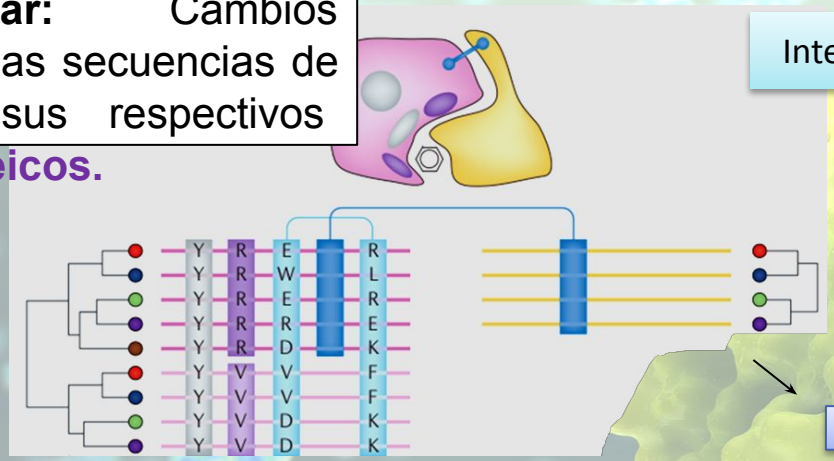
90's



CO-EVOLUCIÓN MOLECULAR

Co-evolución: Interdependecia entre los cambios evolutivos de dos o más **entidades biológicas**.

Nivel molecular: Cambios coordinados en las secuencias de los **genes** y sus respectivos **productos proteicos.**



Zonas altamente conservadas

SDPs

Intra-Proteína

Inter-Proteína

Co-evolución Molecular

Niveles

Inter-residuo

Inter-proteína

Híbrido

(De Juan, 2013).

SPECIFICITY-DETERMINANT POSITIONS (SDPs)

- Grupo de residuos que mutan coordinadamente en el contexto de la divergencia de una subfamilia de proteínas
- Determinantes en el reconocimiento de ligandos.
- Agregados tridimensionales en los sitios activos o sitios de unión.

SDPs

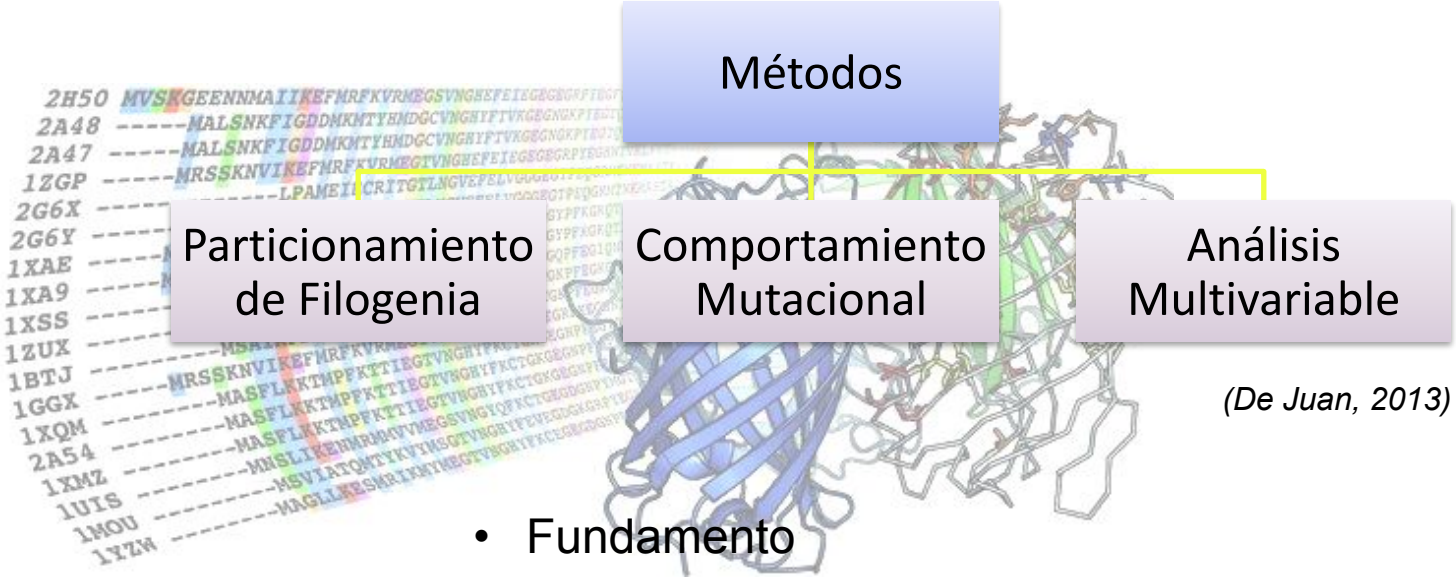
(De Juan, 2013).

Interacciones Proteína-Ligando:

- Complementariedad geométrica y electrostática
- Ligandos pequeños □ Cavidades
- Prot-Prot.□ Superficies amplias y accecibles

- Heterogeneidad sitios funcionales

MÉTODOS DE DETERMINACIÓN DE SDPs



(De Juan, 2013)

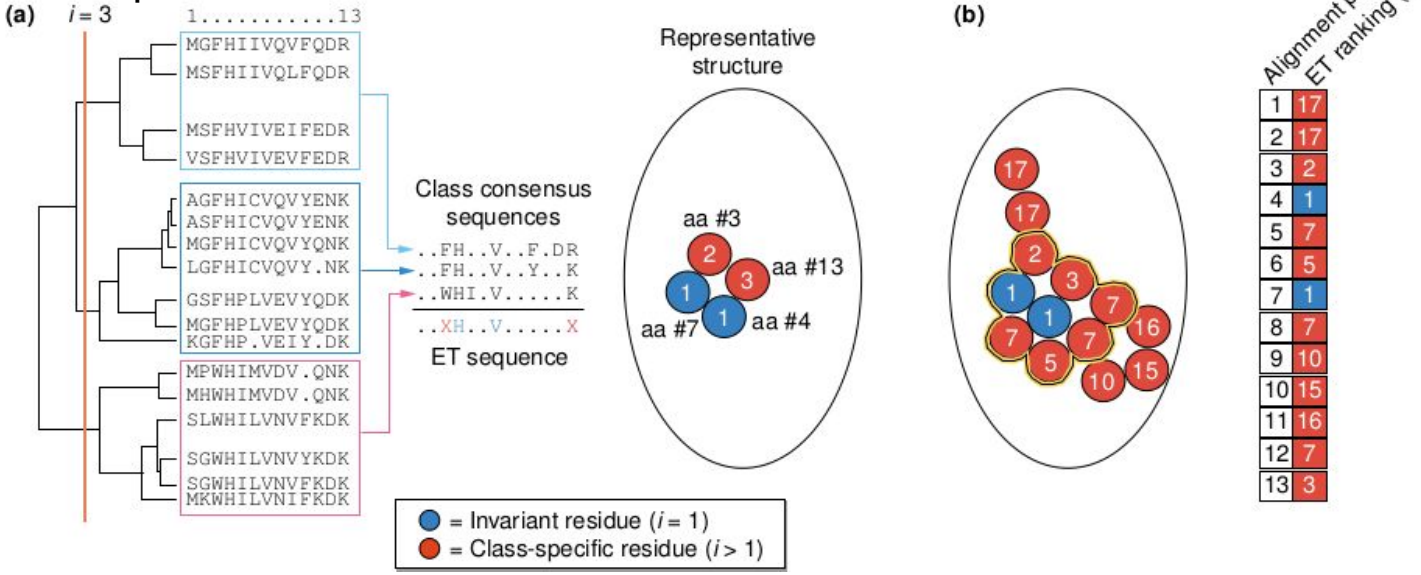
- Fundamento
- Discusión de aplicaciones
- Herramientas bioinformáticas disponibles

1. PARTICIONAMIENTO DE FILOGENIA

Evolutionary Tracing (ET)

(1) Sitios funcionales evolucionan conservando una arquitectura.

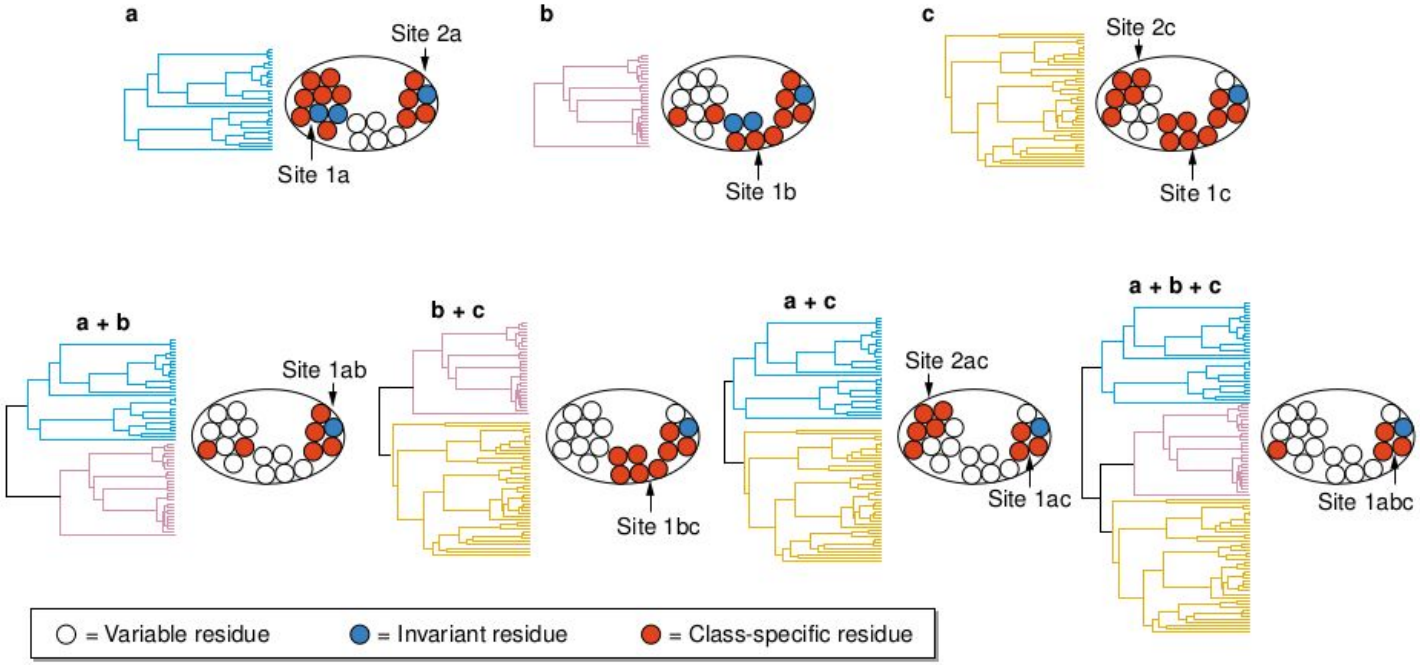
(2) Identidad de los árboles filogenéticos de las secuencias se aproximan a la clasificación funcional.



(Lichtarge, 2002)

1. PARTICIONAMIENTO DE FILOGENIA

Evolutionary Tracing (ET)



ET puede ser usado como filtro funcional.

(Lichtarge,2002)

1. PARTICIONAMIENTO DE FILOGENIA

ET no requiere determinar particionamiento óptimo de MSA.

Métodos optimizados consideran determinación de particionamiento óptimo mediante cálculo de entropía para cada partición.

ET-BASED METHODS

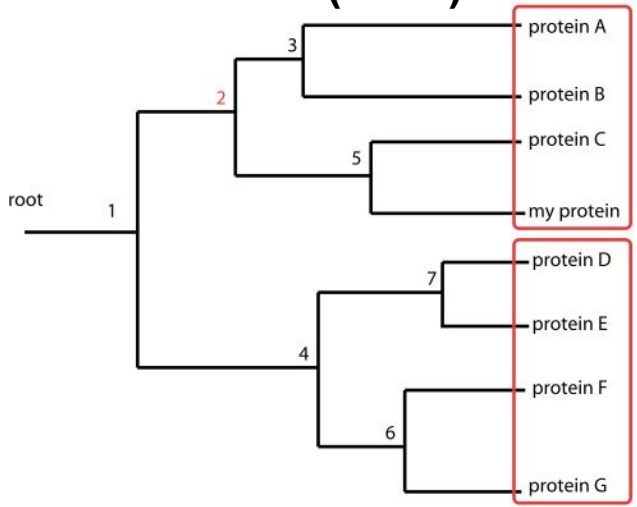
Russell (2000)

S-Method (2003)

rvET (2004)

SDPsite (2009)

Real-Value ET (rvET)



(Lichtarge,2004)

integer-valued ET

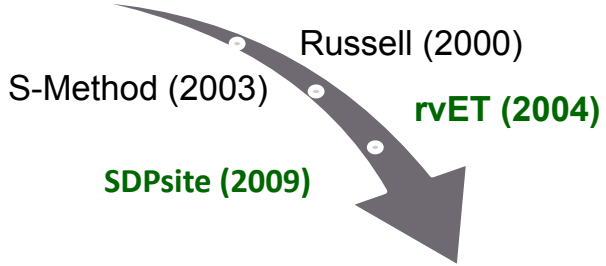
$$r_i = 1 + \sum_{n=1}^{N-1} \times \begin{cases} 0 & \text{if position } i \text{ conserved within each group } g \\ 1 & \text{otherwise} \end{cases}$$

1. PARTICIONAMIENTO DE FILOGENIA

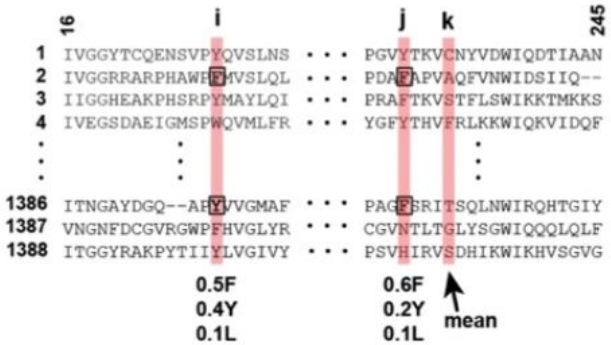
ET no requiere determinar particionamiento óptimo de MSA.

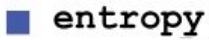
Métodos optimizados consideran determinación de particionamiento óptimo mediante cálculo de entropía para cada partición.

ET-BASED METHODS



Real-Value ET (rvET)





$$s_i = -\sum_{a=1}^{20} f_{ia} \ln f_{ia}$$

0 para posición totalmente conservada

máxima cuando cada aa se encuentra igualmente representado.

(Lichtarge,2004)

1. PARTICIONAMIENTO DE FILOGENIA

ET no requiere determinar particionamiento óptimo de MSA.

Métodos optimizados consideran determinación de particionamiento óptimo mediante cálculo de entropía para cada partición.

ET-BASED METHODS

Russell (2000)

S-Method (2003)

rvET (2004)

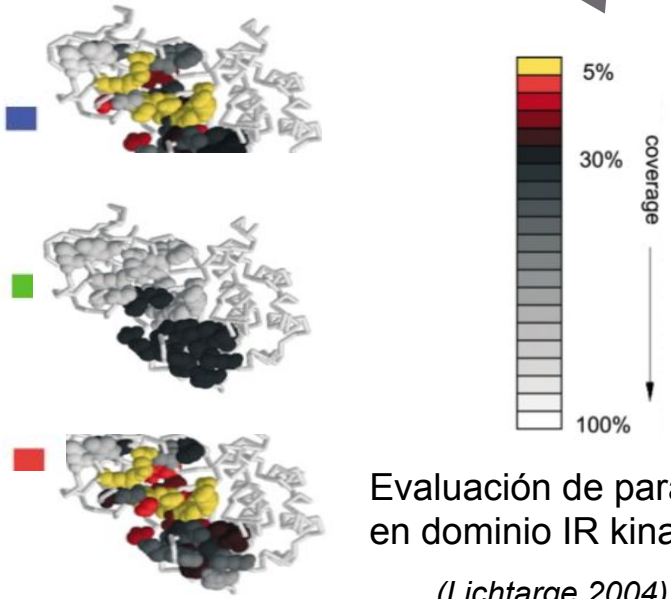
SDPsite (2009)

Real-Value ET (rvET)

real-valued ET

$$\rho_i = 1 + \sum_{n=1}^{N-1} \frac{1}{n} \sum_{g=1}^{n} \left(- \sum_{a=1}^{20} f_{ia}^{g} \ln f_{ia}^{g} \right)$$

Herramienta involucra también otros parámetros asociados a asignación de pesos a secuencias, entre otros.



Evaluación de parámetros en dominio IR kinasa (1IRK)

(Lichtarge,2004)

2. COMPORTAMIENTO MUTACIONAL

Mutational Behaviour (MB)

(1) Posiciones que mejor reflejan historia evolutiva de la familia se asocian a especificidad.

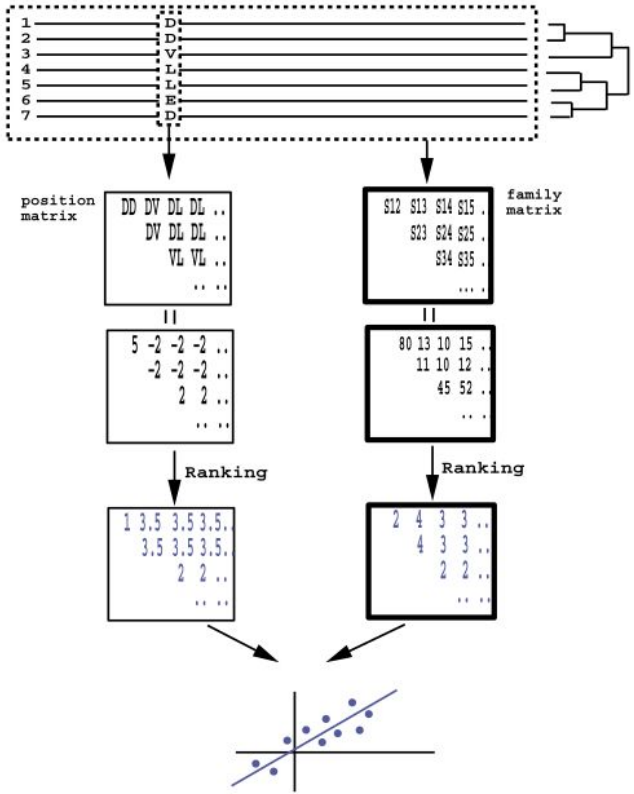
(2) La variación entre estas posiciones tiende a estar correlacionada.

Comparación del patrón de mutación de cada posición del MSA respecto a la variación total de las secuencias.

Spearman Rank-Order Correlation Coefficient

$$r = \frac{\sum_i (R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\sum_i (R_i - \bar{R})^2}\sqrt{\sum_i (S_i - \bar{S})^2}}$$

Xset. Integrado en JDet (2012)



* Matriz de cambio

* Función de puntaje

(Valencia,2003)

3. ANÁLISIS MULTIVARIABLE

(1) Detección de principales fuentes de variabilidad en un MSA, y las posiciones asociadas, mediante **análisis de componentes principales (PCA).**

PCA-BASED METHODS

SeqSpace (1995)

S3Det (2003)

Jdet (2012)

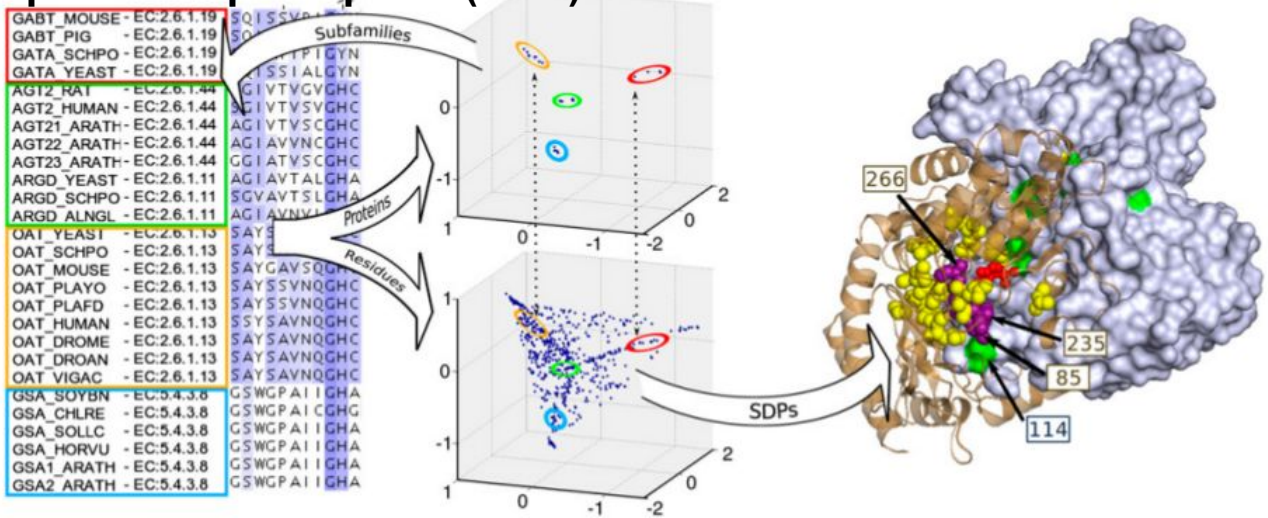
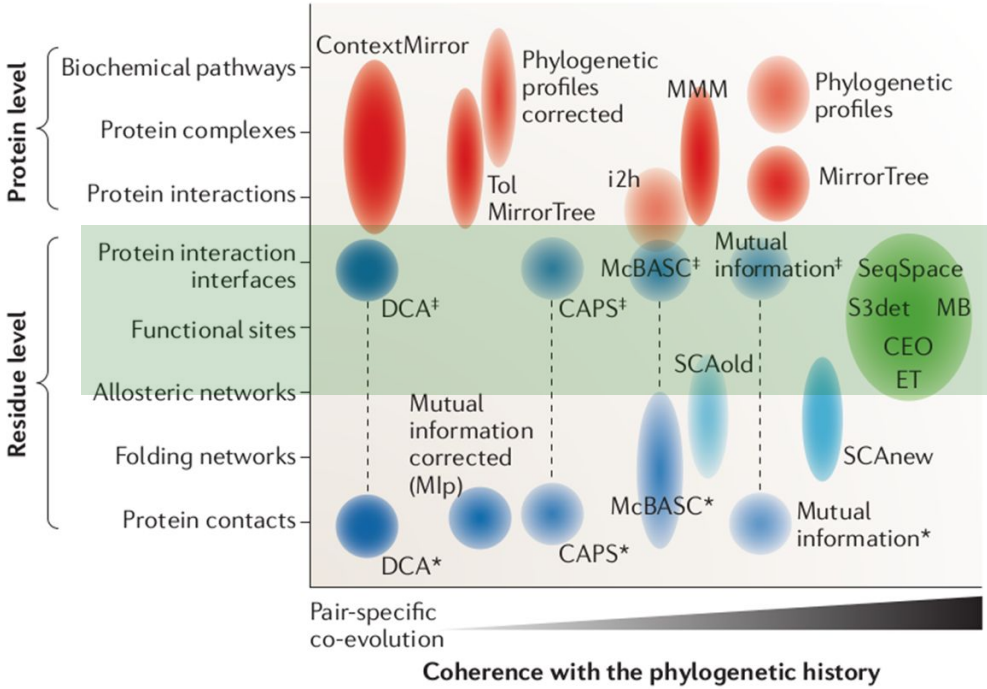


Fig. 2. Workflow implemented to simultaneously detect the protein subfamilies and those residues responsible for such segregation (SDPs). This process is depicted for the class III aminotransferase family (Pfam PF00202). The homodimeric structure of the human ornithine aminotransferase (PDB 1oat) bound to Pyridoxal-5′-phosphate (in red spheres) is represented. The two subunits of the complex are shown as brown cartoon representation and with a gray surface. SDPs are highlighted in a yellow/violet spacefill and with a green surface. The figure was generated with Pymol (pymol.sourceforge.com).

(Rausell, 2010)

APLICACIONES



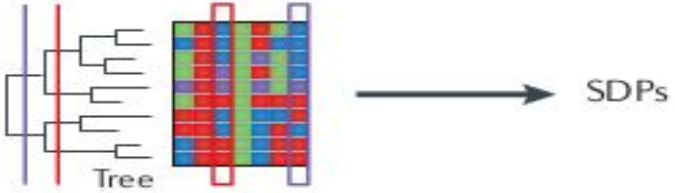
MÉTODOS DETERMINACIÓN SDPs

(De Juan, 2013).

APLICACIONES

ET:

- Sólo considera aa idénticos como sitios conservados.
- Requiere estructura para contextualizar sitios funcionales.
- Requiere análisis manual de las posiciones detectadas, las cuales se pueden asociar a diferentes actividades (unión de ligando, alosterismo, rol estructural).
- No requiere detección previa de sub-familias.



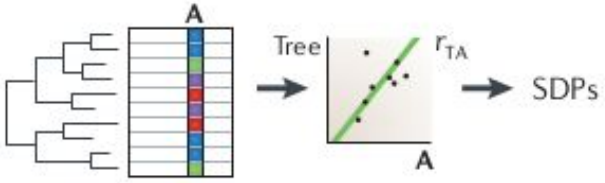
APLICACIONES

MB:

- Detección de claros cambios en las propiedades de los aa y su influencia en la divergencia de las secuencias.

- Se ha hallado particularmente útil en la detección de SDPs altamente correlacionados con la divergencia de una familia de proteínas.

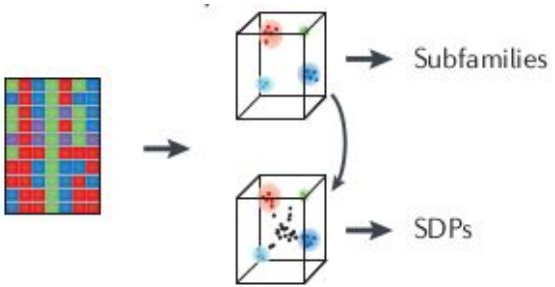
- Tienden a corresponder a interfaces de unión de ligandos o proteínas.



APLICACIONES

PCA:

- Permite representación de asociación entre posiciones para características de cada sub-familia en espacio multidimensional.
- Detección de SDPs sencilla.
- Requiere inspección manual de los espacios.



- Ha mostrado utilidad en la guía de estudios experimentales.

(De Juan, 2013)

HERRAMIENTAS BIOINFORMÁTICAS DISPONIBLES

SDPs			
Evolutionary trace	SDPs	Ligand and protein interaction specificity	Evolutionary Trace Server (http://mammoth.bcm.tmc.edu/ETserver.html)
SDPsite	SDPs and subfamilies	Ligand and protein interaction specificity	SDPsite (http://bioinf.fbb.msu.ru/SDPsite/index.jsp)
Mutational behaviour	SDPs	Ligand and protein interaction specificity	Jdet (http://csbg.cnb.csic.es/JDet/)
S3det	SDPs and subfamilies	Ligand and protein interaction specificity	

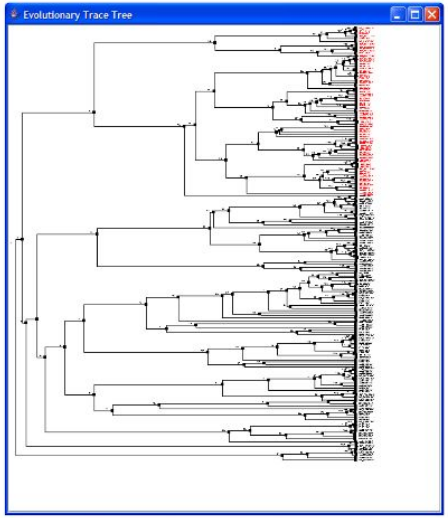
(De Juan, 2013)

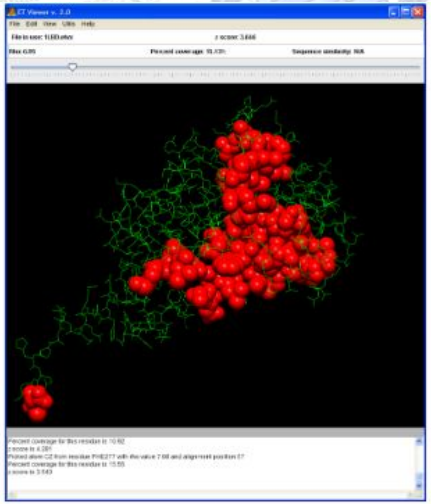
HERRAMIENTAS BIOINFORMÁTICAS DISPONIBLES

ET Viewer

ET can be performed from a variety of starting points. Typically, a new trace is run by providing ET with a structure file (in PDB format) that is representative of the protein of interest. A trace can also be initiated with an existing multiple sequence alignment (GCG format), phylogenetic tree (NHX format), list of sequence names from BLAST, or a list of sequences in FASTA format. In all of these cases it is most useful to provide a PDB structure since the power of ET is visualization of ET residues on a 3-dimensional structure. In fact, when running ET via the ET Wizard, you are required to provide a structure and/or PDB code to successfully complete ET.

(Lichtarge,2006)

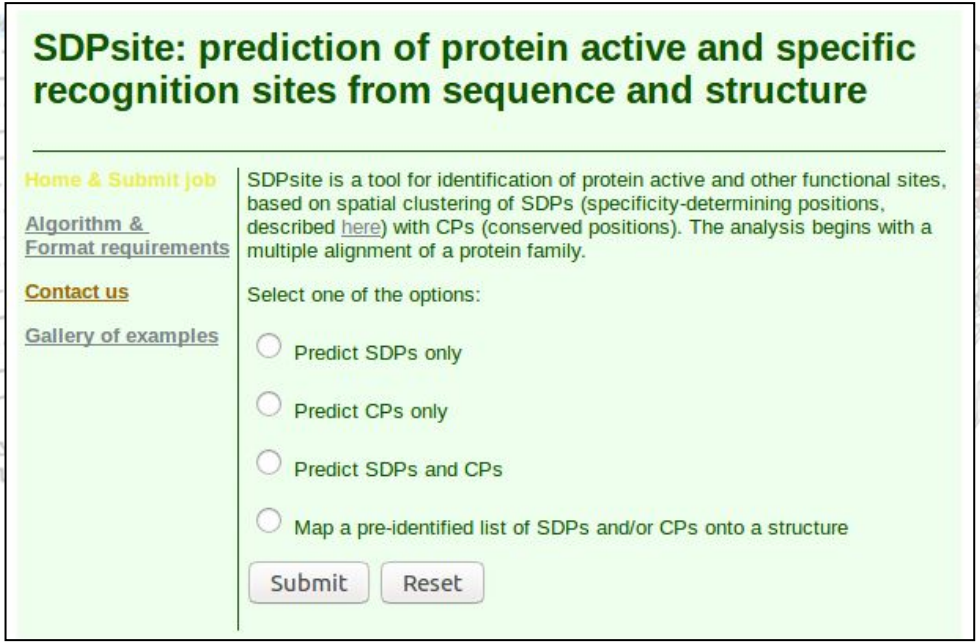




Evolutionary Trace Viewer 2.0 Manual April 5, 2007

HERRAMIENTAS BIOINFORMÁTICAS DISPONIBLES

SDPsite



(Russell,2009)

HERRAMIENTAS BIOINFORMÁTICAS DISPONIBLES

SDPsite

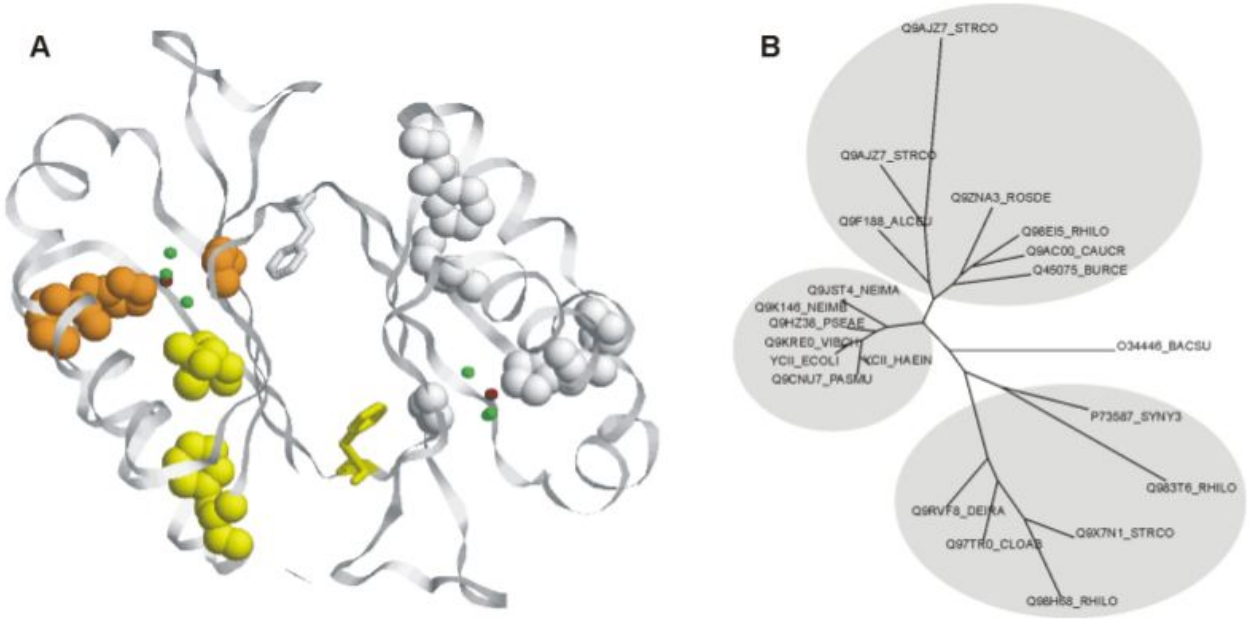


Figure 3
A. Structure of HI0828 from *Haemophilus influenzae* (1 mwq). SDPs are marked yellow, CPs are marked orange, best cluster is shown in spheres. Cl ions are shown in green, Zn ions in brown. **B. Phylogenetic tree of the YCII-related domain (PF03795) family.** The predicted specificity groups are shown as gray ovals.

(Russell,2009)

HERRAMIENTAS BIOINFORMÁTICAS DISPONIBLES

Jdet

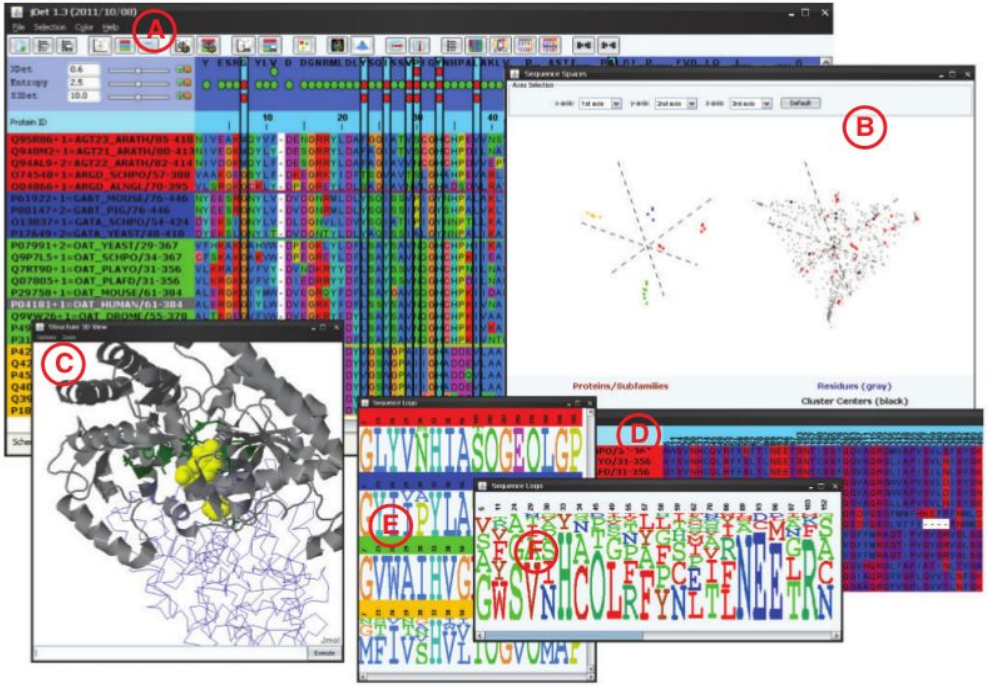
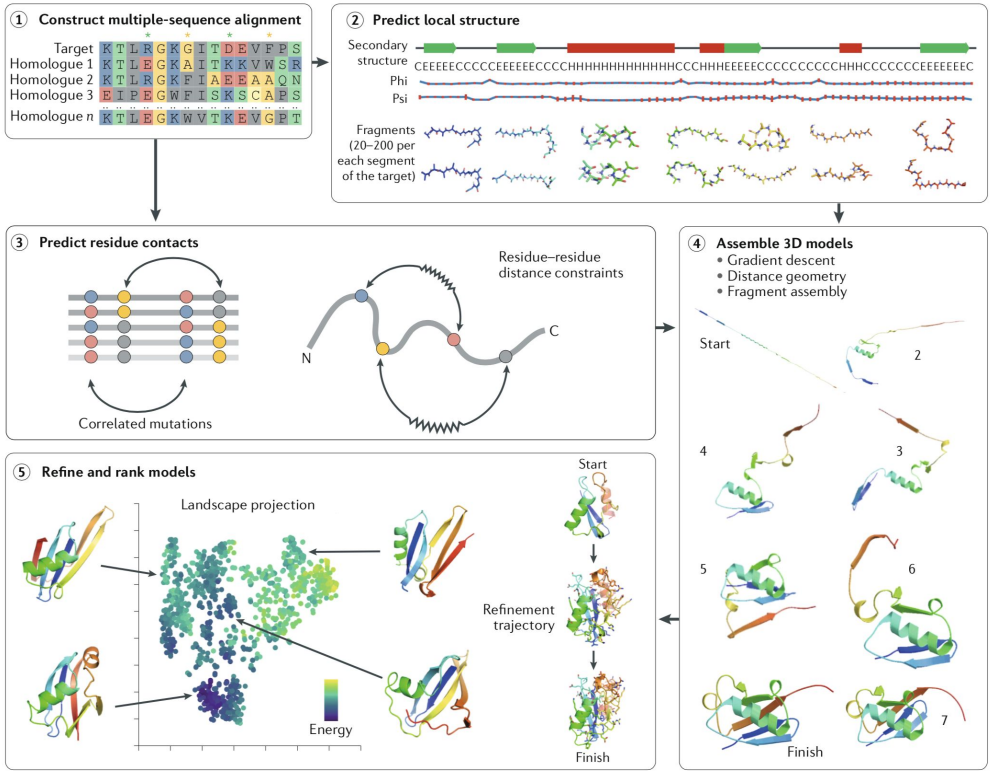


Fig. 1. Screenshots of the JDet interface. (**A**) Main window with an alignment and the results of three methods loaded into it. The subfamilies automatically detected are shown with different colors on the left. (**B**) 3D interactive views of the protein and residue spaces generated by S3Det. (**C**) Jmol window showing the 3D structure of one of the proteins of the alignment. Panels (**A**)–(**C**) are communicated with each other: proteins or residues selected in any of them become selected in the other two. (**D**) Subalignment showing only the currently selected proteins and positions, and with amino acids colored according to their polarity. (**E**) Sequence logos for the selected positions/proteins calculated for each subfamily independently. (**F**) Global sequence logos for the selected positions/proteins, and amino acids colored according with their polarity.

(Pazos,2011)

Cálculo de Estructuras 3D Desde Secuencia

Free Structure Prediction



Kuhlman, Brian, and Philip Bradley. 2019. “Advances in Protein Structure Prediction and Design.” *Nature Reviews Molecular Cell Biology*, August, 1–17. https://doi.org/10.1038/s41580-019-0163-x.

AlphaFold

AlphaFold method (Jumper *et al*, 2021) by the Google DeepMind team, and the analogous RoseTTAfold approach by the Baker laboratory at the University of Washington, Seattle, USA (Baek *et al*, 2021), a few months later, explain the background behind Deep Learning (DL) algorithms that made this scientific milestone possible.

The DeepMind team, collaborating with the European Bioinformatics Institute (EBI), then applied AlphaFold to whole genomes, including human, mouse, *Saccharomyces,* and *E. coli* (Tunyasuvunakool *et al*, 2021). https://alphafold.ebi.ac.uk

The SwissProt have integrated the AlphaFold prediction into the website.

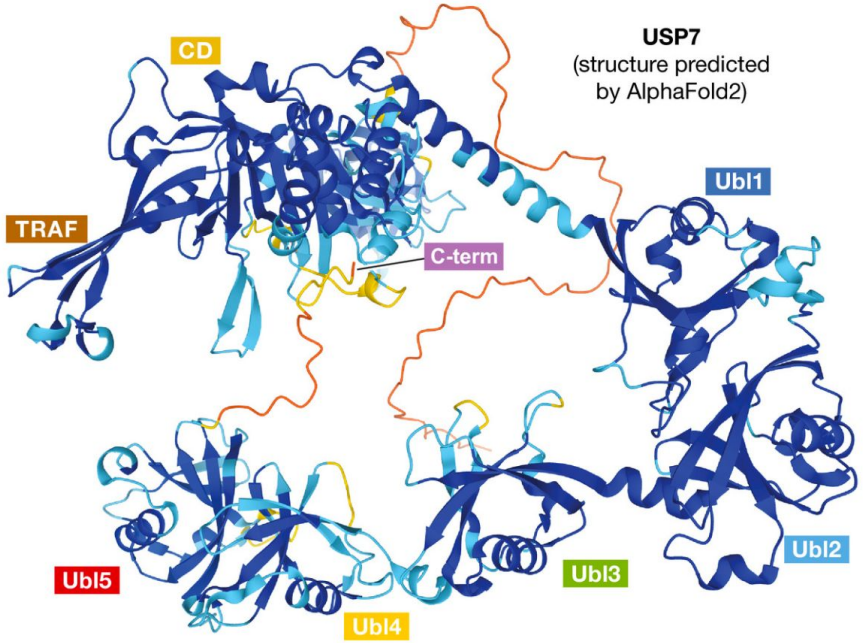
Baek, Minkyung, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, et al. 2021. “Accurate Prediction of Protein Structures and Interactions Using a Three-Track Neural Network.” *Science* 373 (6557): 871–76. https://doi.org/10.1126/science.abj8754.

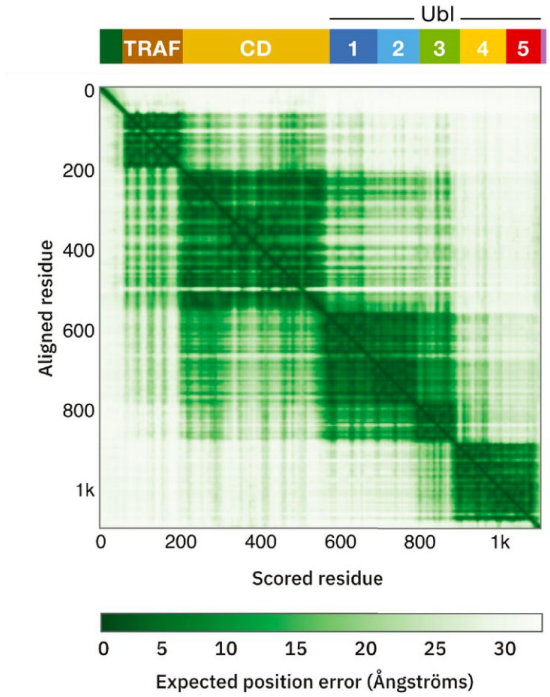
Jumper, John, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, et al. 2021. “Highly Accurate Protein Structure Prediction with AlphaFold.” *Nature* 596 (7873): 583–89. https://doi.org/10.1038/s41586-021-03819-2.

Tunyasuvunakool, Kathryn, Jonas Adler, Zachary Wu, Tim Green, Michal Zielinski, Augustin Žídek, Alex Bridgland, et al. 2021. “Highly Accurate Protein Structure Prediction for the Human Proteome.” *Nature* 596 (7873): 590–96. https://doi.org/10.1038/s41586-021-03828-1.

AlphaFold

https://alphafold.ebi.ac.uk





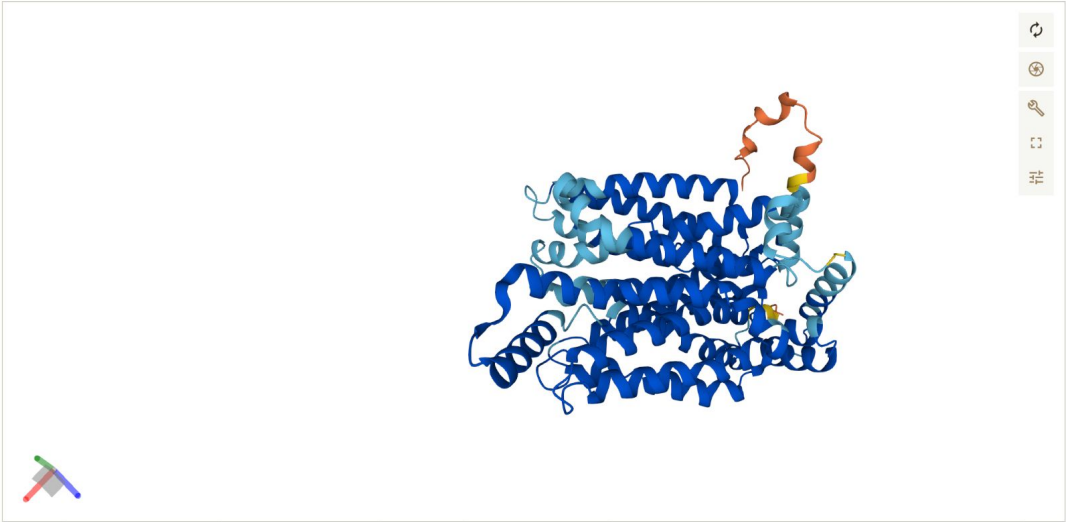
AlphaFold

https://www.uniprot.org/uniprot/P11166#structure

Model Confidence:

- Very high (pLDDT > 90)
- Confident (90 > pLDDT > 70)
- Low (70 > pLDDT > 50)
- Very low (pLDDT < 50)

AlphaFold produces a per-residue confidence score (pLDDT) between 0 and 100. Some regions with low pLDDT may be unstructured in isolation.



SOURCE -- Select --	IDENTIFIER	METHOD -- Select --	RESOLUTION	CHAIN	POSITIONS	LINKS
PDB	4PYP	X-ray	3.17 Å	A	1-492	PDBe · RCSB-PDB · PDBj · PDBsum
PDB	5EQG	X-ray	2.90 Å	A	1-492	PDBe · RCSB-PDB · PDBj · PDBsum
PDB	5EQH	X-ray	2.99 Å	A	1-492	PDBe · RCSB-PDB · PDBj · PDBsum
PDB	5EQI	X-ray	3.00 Å	A	1-492	PDBe · RCSB-PDB · PDBj · PDBsum
PDB	6THA	X-ray	2.40 Å	A	1-492	PDBe · RCSB-PDB · PDBj · PDBsum
AlphaFold	AF-P11166-F1	Predicted			1-492	AlphaFold

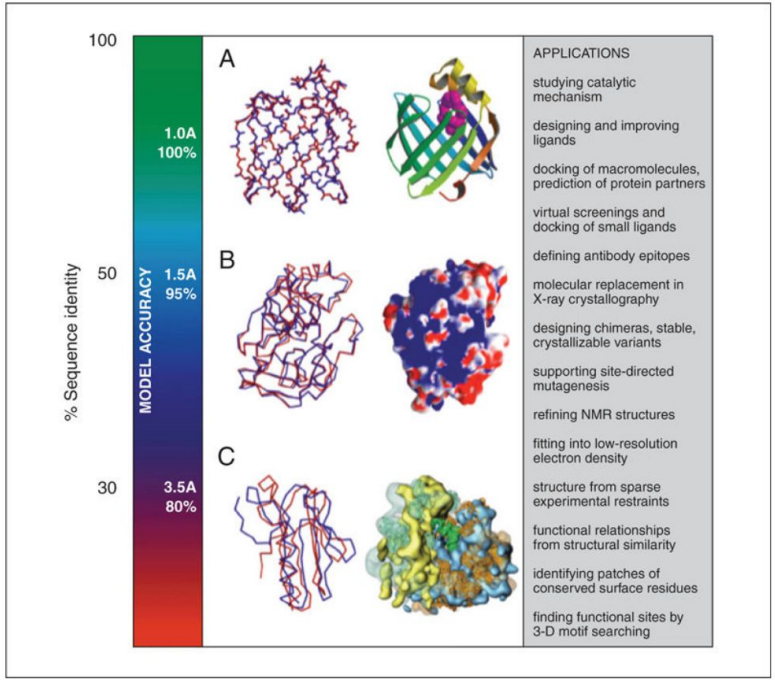
AlphaFold

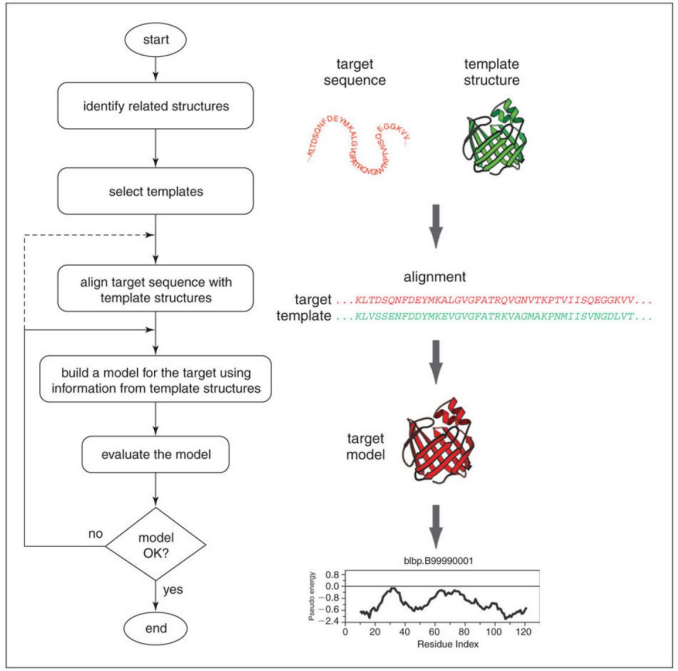
https://github.com/deepmind/alphafold

This idea of co-evolution is relatively simple: If two residues are close in space and interact with each other—even if they are far apart in the amino acid sequence—they will stick together during evolution to preserve structure and function. T

Modelamiento Comparativo

Comparative Modelling





Webb, Benjamin, and Andrej Sali. 2016. “Comparative Protein Structure Modeling Using MODELLER.” *Current Protocols in Bioinformatics* 54 (1): 5.6.1-5.6.37. https://doi.org/10.1002/cpbi.3.

Modeller Tips

- Find and analyze templates availables:
  - BLAST search
  - Modeller search
  - Conformations, Resolutions, Ligands
- Build and analyze alignment:
  - Coverture
  - PDB postions
  - Gaps
- Models build and Applications
  - One or many models
  - Refinement Process
  - Molecular Dynamic Exploration

Comparative Modelling

```
Sequence identity comparison (ID_TABLE):

   Diagonal        ... number of residues;
   Upper triangle  ... number of identical residues;
   Lower triangle  ... % sequence identity, id/min(length).

          1b8pA @1 1bdmA @1 1civA @2 5mdhA @2 7mdhA @2 1smkA @2
1b8pA @1      327     194     147     151     153      49
1bdmA @1       61     318     152     167     155      56
1civA @2       45      48     374     139     304      53
5mdhA @2       46      53      42     333     139      57
7mdhA @2       47      49      87      42     351      48
1smkA @2       16      18      17      18      15     313

Weighted pair-group average clustering based on a distance matrix:

                                        ---------------------- 1b8pA @1.9   39.0000
                                        |
                                 -------|--------------------- 1bdmA @1.8   50.5000
                                 |
                            -----|------------------------------ 5mdhA @2.4   55.3750
                            |
                            |                                  --- 1civA @2.8   13.0000
                            |                                  |
      ------------------------------------------------------------ 7mdhA @2.4   83.2500
      |
    -------------------------------------------------------------- 1smkA @2.5

    +----+----+----+----+----+----+----+----+----+----+----+----+
  86.0600   73.4150   60.7700   48.1250   35.4800   22.8350   10.1900
      79.7375   67.0925   54.4475   41.8025   29.1575   16.5125
```

Figure 5.6.6.
Excerpts from the log file `compare.log`.

```
 _aln.pos        10        20        30        40        50        60
ThdmA    MKAPVRVAVTGAAGQIGYSLLFRIAAGEMLGKDQPVILQLLEIPQAMKALEGVVMELEDCAFPLLAGL
TvLDH    MSEAAHVLITGAAGQIGYILSHWIASGELYG-DRQVYLHLLDIPPAMNRLTALTMELEDCAFPHLAGF
_consrvd *     *  ********* *   ** **  * *  * * ** ** **  *    ********* ***

_aln.p   70        80        90       100       110       120       130
ThdmA    EATDDPDVAFKDADYALLVGAAPRL --------QVNGKIFTEQGRALAEVAKKDVKVLVVGNPANTN
TvLDH    VATTDPKAAFKDIDCAFLVASMPLKPGQVRADLISSNSVIFKNTGEYLSKWAKPSVKVLVIGNPDNTN
_consrvd  ** **  **** * * **   *              *  **   *  *   **  ***** *** ***

_aln.pos  140       150       160       170       180       190       200
lbdmA    ALIAYKNAPGLNPRNFTAMTRLDHNRAKAQLAKKTGTGVDRIRRMTVWGNHSSIMFPDLFHAEVD---
TvLDH    CEIAMLHAKNLKPENFSSLSMLDQNRAYYEVASKLGVDVKDVHDIIVWGNHGESMVADLTQATFTKEG
_consrvd   **   *  * * **     ** ***    * * *  *       *****   *  **  *

_aln.pos    210       220       230       240       250       260       270
ThdmA    -GRPALELVDMEWYEKVFIPTVAQRGAAIIQARGASSAASAANAAIEHIRDWALGTPEGDWVSMAVPS
TvLDH    KTQKVVDVLDHDYVFDTFFKKIGHRAWDILEHRGFTSAASPTKAAIQHMKAWLFGTAPGEVLSMGIPV
_consrvd          *       *       *   *   **  ****   *** *    * **  *   **  *

_aln.pos      280       290       300       310       320       330
ThdmA    Q--GEYGIPEGIVYSFPVTAK-DGAYRVVEGLEINEFARKRMEITAQELLDEMEQVKAL--GLI
TvLDH    PEGNPYGIKPGVVFSFPCNVDKEGKIHVVEGFKVNDWLREKLDFTEKDLFHEKEIALNHLAQGG
_consrvd      ***  * * ***      *   ****   *   *     *   *  * *
```

Figure 5.6.8.
The alignment between sequences `TvLDH` and `1bdmA`, in the MODELLER PAP format. File `TvLDH-1bmdA.pap`.

Webb, Benjamin, and Andrej Sali. 2016. “Comparative Protein Structure Modeling Using MODELLER.” *Current Protocols in Bioinformatics* 54 (1): 5.6.1-5.6.37. https://doi.org/10.1002/cpbi.3.

Lecture 03: Proteins Modelling

Dr. Alexis Salas Burgos
Marzo 2022