Protein Structures

Vinay Kakkar

December 4, 2022

Contents

Abstract

Three-dimensional structures are created from sequences of amino acids in polypeptide chains folds that are generated from linear chains. The folded domains can serve as modules for building up large assemblies such as a muscle fiber but more importantly, they can provide specific catalytic or binding sites, as found in enzymes or proteins. We look at the foundations of a protein such as the building blocks. Using technologies in order to read these sequences of amino acids into information that helps determine what the function of the structure will have. Using our knowledge of protein structures and bioinformatics, we can achieve the goal of predicting functions of proteins from a large set of data provided by experiments such as, DNA microarray technology and Two-dimensional Gel electrophoresis or Chromatography. With new technologies such as AlphaFold, we can even predict the 3D structure of a Protein using the amino acid sequence alone.

1 Introduction

Amino acids are molecules that when combined together it forms proteins. All of the 20 amino acids, see table 1 have in common a central carbon atom which are attached a hydrogen atom, an amino group and a carboxyl group. What distinguishes one amino acid from another is the side chain attached to the central carbon atom through its fourth valence [BT98].

Amino acid	Three-letter code	One-letter code
Glycine	Gly	G
Alanine	Ala	A
Valine	Val	V
Leucine	Leu	L
Isoleucine	Ile	I
Proline	Pro	P
Phenylalanine	Phe	F
Methionine	Met	M
Tryptophan	Trp	W
Cysteine	Cys	C
Asparagine	Asn	N
Glutamine	Gln	Q
Serine	Ser	S
Threonine	Thr	T
Tyrosine	Tyr	Y
Aspartic acid	Asp	D
Glutamic acid	Glu	E
Histidine	His	H
Lysine	Lys	K
Arginine	Arg	R

Table 1: The 20 amino acids. The amino acid name, the three-letter code, and the one-letter code are given. The Amino acids are split up into Nonpolar, Polar, Acidic and Basic respectfully

Proteins are responsible of catalysing most the chemical reactions in cells. They can function as enzymes catalysing a wide variety of reactions important for life and thus also important for the structure of living systems such as proteins involved in the cytoskeleton. The size of protein can vary [Zve08].

Definition 1.1 (Catalysing) *Catalysing is to make a chemical reaction happen or happen more quickly by acting as a catalyst.*

Definition 1.2 (Cytoskeleton) *A dynamic network of interlinking protein filaments present in the cytoplasm of all cells [Zve08].*

We can analyse a DNA sequence of a gene to retrieve the amino acid sequence of the protein product, using the fact that protiens are a built up off amino acids.Leaving a position where we can help deduce the likely properties of unknown proteins, whilst at the same time including their functions and structures. Knowing the relationship between a proteins structure and its function provides a better understanding of how the protein works ti better understand this we can conduct experiments to explore how modifying the structure will affect the function. The use of bioinformatics aids this process whilst also providing computer modelling for these interactions [Zve08].

2 Protein Structure

2.1 Primary, Secondary, Tertiary and Quaternary Structure

Primary Structure

The primary structure of a peptide or protein is the linear sequence of its amino acids. It is read and written from the amino-terminal to the carboxyl-terminal end. Where each amino acid is connected to the next by a peptide bond. Primary strucutre sequence it can interact with one another to form secondary structures [oV].

Secondary Structure

The secondary structure refers to the local arrangement of a peptide chain. Where several common secondary structures have been identified in proteins [oV].

Tertiary Structure

Tertiary structure is a three-dimensional structure of a protein the formation is built up of bonds and interactions that serve to change the shape of the overall protein. Finaly the folding that we end up with for a given polypeptide is the tertiary structure [God22].

Quaternary Structure

Quaternary structure of a protein is the built up of several protein chains/subunits. Each of the subunits has its own primary, secondary, and tertiary structure. The subunits are held together by hydrogen bonds and van der Waals forces between nonpolar side chains [OR15].

Definition 2.1 (Van Der Waals) *A relatively weak electric force that attract neutral molecules that collide with or pass very close to each other [Bou18].*

Protein	Number of Subunits	Function
Alcohol dehydrogenase	4	Enzymatic reaction in fermentation
Aldolase	4	Enzymatic reaction in glycolysis
Fumarase	4	Enzymatic reaction in citric acid cycle
Hemoglobin	14	Oxygen transport in blood
Insulin	2	6344

Table 2: Examples of Proteins Having Quaternary Structure [OR15].

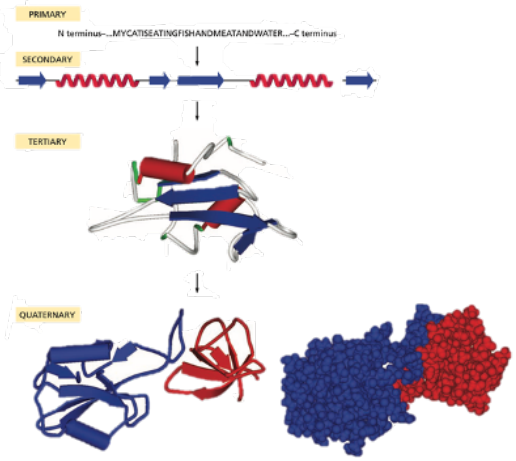


Figure 1: From the sequence alone, the primary structure to secondary structure, to tertiary structure(3D), to finally quaternary structure found when several tertiary structures form a multisubunit complex [Zve08].

Considering Protein structure on several different levels

The fold of the protein plays part in determining the way the protein will function, and also whether it will function correctly so it is important to understand these folds. Which we can use to help us for example predicting the fold of a protein from its sequence. Looking at Protein structures on different levels we need to consider the analysis of protein structure by experimental techniques such as X-ray crystallography, nuclear magnetic resonance and RNAseq which show that proteins adopt distinct structural elements [Zve08].

Amino Acids

When looking at a primary structure of a protein the sequence of amino acids will build up the linear protein chain. This linear chain is often called a polypeptide chain [Zve08].

Amino acids are different to each other due to their side chains and due to this the functional properties various different proteins are different. Each type of amino acid has specific chemical physical properties determined by the structure and chemical properties of its side chain. They can, however, be classified into overlapping groups that share some common physical and chemical properties [Zve08].

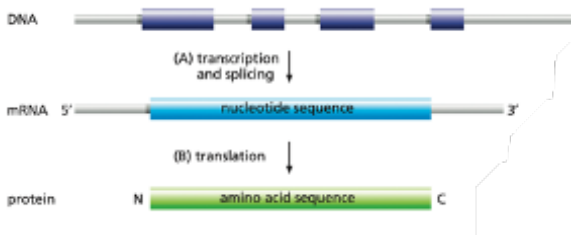


Figure 2: The relation of DNA coding-strand sequence to mRNA sequence to protein sequence. The exons (purple boxes) of the DNA are transcribed into mRNA which, using other molecules directs the protein sequence [Zve08].

Bioinformatic Difficulties with Predictions on Proteins

It is difficult to define the percise ends of the helices(Secondary structure of proteins is made up of a-helices and b-strands) for structures found in globular proteins that are not perfectly regular. Making it one step more difficult when trying to predict these structures [Zve08].

To Note:

- Several different types of b-sheet are found in protein structures.
- Turns, hairpins, and loops connect helices and strands.
- Any chain between two regular structures is referred to as a loop.
- Mostly a loop will contain a turn (or even several).

In antibody recognition, immunoglobulins employ loops at the edge of a b-sheet. All immunoglobulin structures with the same overall chain fold, but it is the difference at these loops that results in different results. Loops take up one of a limited number of structures called canonical forms. This type of classification is another reason why trying to predict both the structure and function of the protein is difficult [Zve08].

Definition 2.2 (Immunoglobulin) *Immunoglobulins are heterodimeric proteins composed of two heavy and two light chains. Types of white blood cells that helps the body fight infection [SC10].*

2.2 Compact Structures

Protein Folds

Protein chains by themselfes no major function. Only does so once the chain has folded up into a tertiary or quaternary structure [Zve08].

For example:

Proteins enzymes bind to other molecules (ligands) and catalyze their biochemical reactions, or they can influence proteins activity, or regulate gene expression by binding to DNA. In other cases proteins have a purely structural function for example, making up the fabric of the cell. Or for example, proteins acting as chemical messengers are released, from cells which, influence the behavior of other cells by acting on another class of proteins, known as receptors [Zve08].

The tertiary structure is defined by the path of its polypeptide chain in the tertiary structure of a protein where each domain folds independently of the others [Zve08].

Bioinformatics with Protein Folds

Bioinformatics more intrested in the sequences and structures of different domains over whole proteins where a domian can vary from 50 to around 350 amino acids in length [Zve08].

There is a limited number of ways secondary structures fold into domains.

For example we have had proteins that seem to be completely unrelated in terms of sequence are found to have the same fold giving the idea that there may be finite number of folds in nature.

As proteins can fold into a similar structure even if their sequences are not very similar means that we can use bioinformatics tools to model structures of various proteins on similar folds [Zve08].

3 Large Scale Experssion

Gene expression begins when genes are transcribed into messenger RNAs, which are then translated to produce proteins.

Total gene expression in cultured cells or a tissue sample can be detected in three main ways:

1. DNA microarray technology.
2. Two-dimensional Gel electrophoresis or Chromatography.
3. RNAseq

With the first two cases they produce enormous amounts of raw data [Zve08] due to this, many proteins currently evade high-resolution structure determination. Structural mass spectrometry is a powerful approach that better then the first two methods mentioned above, by having nearly a unlimited size constraint, and speed. Although the data provided by mass spectrometry is vauge for full high-resolution structure elucidation, structural mass spectrometry can be used to examine

the size, solvent accessibility, and topography of proteins [LLV18] [LZG20]. Many mass spectrometry techniques exist that can elucidate elements of protein tertiary and quaternary structure [BL22].

We can have computational methods that aid experimental technique with the goal to elucidate protein structures [SL20] [LWL+20]. Software packages can be used to combine data with advanced structure sampling and scoring techniques. Computational tools for protein structure modeling, include the Rosetta software suite [LWL+20] [ALFJ+17], I-TASSER [YYR+15], Phyre2 [KMY+15], Integrative Modeling Platform [RLW+12], HADDOCK [DBB03], and MODELLER [EWMR+06] [BL22].

3.1 Large Scale Gene Expression

Genome DNA microarray experiments produce large amount of data can be computationaly heavy on where methods can yield alternative conclusions from inceasing the computational effort.

The goal of these experiments is to determine biological or functional meaning from the lists of genes, either by:

1. Identify critical genes that are responsible for a biological effect.
2. Find patterns within the genes that point to an underlying biological process.

[Zve08]

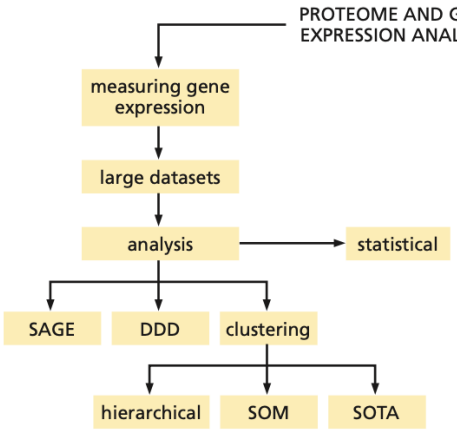


Figure 3: Describing Common experimental aspects of gene expression and of the analysis of the resulting data [Zve08].

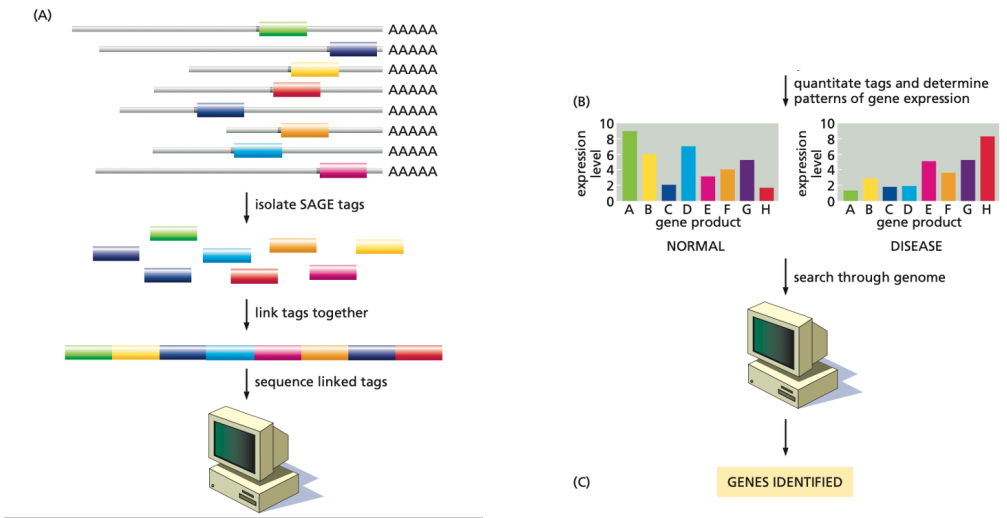
Serial analysis of gene expression

Serial analysis of gene expression is the alternative compared to microarrays when trying to investigate patterns of gene expression.

A short sequence contains enough information to uniquely identify a gene. The sequence tags from the total cellular RNA can be linked together to form long DNA molecules. The total number of times a particular tag is observed the concatemers approximates the expression level of the corresponding gene. The data produced by SAGE include a list of the tags with their corresponding counts, providing a digital output of cellular gene expression. Which allow the user to specify which organ is to be investigated. Libraries consisting of gene lists organized by the various types of tissues or cell lines are provided for further choice. The output from SAGE provides the SAGE tag, the UniGene ID, the gene description, and color and letter coded differences in expression levels [Zve08].

Clustered gene expression data

Clustered pattern data obtained from gene expression microarrays/genome bioinformatics can be used as a tool to identify new transcription factors or other cell-regulatory proteins.



(a) (b)

Figure 4: *An outline of the SAGE method for comparing levels of gene expression. (A) Short sequence tags. The sequence tags are isolated and are linked together to produce long DNA molecules that can be cloned and sequenced. (B) Once sequenced, each tag can be calculated, resulting in a value that gives the expression level of the corresponding transcript [Zve08].*

The clustered genes/proteins can be analyzed. Leading to a vast collection of data from many gene/protein expression expeniments being available on the Web [Zve08].

3.2 Large Scale Protein Expression

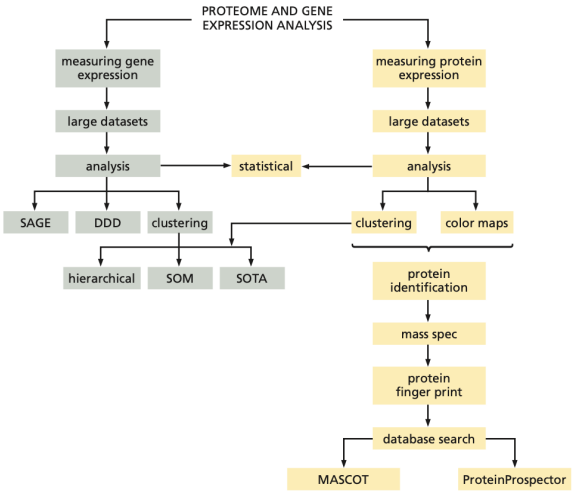


Figure 5: Describing some experimental aspects of protein expression and of the analysis of the resulting data. [Zve08].

For functional protein, mRNAs need to be translated, whilst the protein products can change which influence their function. For this reason we can measure and anlayse different proteins.

There is more proteins than there are genes in a genome. Transcripts can be spliced in various ways to give different mRNAs, providing different protein products, from the same gene. However, proteins that can be modified after translation giving more different protien products.

Protien expressions can vary in a organism depending on the origin and it will also differ between the separate stages of an organism's life cycle and under different environmental conditions.

Definition 3.1 (proteome) *The proteome refers to all the proteins that make up an organism at a specific point in time and under specific conditions.*

It is important to know how protein expression is affected in order to understand how an organism or a cell functions [Zve08].

3.3 RNAseq

Transcriptome is important for revealing the molecular constituents of cells and tissues, interpreting the functional aspects of the genome and, also for understanding development and disease [WGS09].

Many methods deduce and quantify the transcriptome, including hybridizationor sequence-based approaches. For example hybridization-based approaches which involve incubating fluorescently labelled cDNA with microarrays or commercial high-density oligo microarrays [WGS09].

However, these methods have several limitations, such as:

- Dreliance upon existing knowledge about genome sequence.
- Limited dynamic range of detection owing to both background.
- High background levels owing to cross-hybridization [OM06] [RRG07].
- saturation of signals.

Definition 3.2 (transcriptome) *The transcriptome is the complete set of transcripts in a cell, and their quantity, for a specific developmental stage or physiological condition.*

Sequence-based approaches directly determine the cDNA sequence such as, Tag-based methods which include SAGE, CAGE [KKN+06], MPSS [RBL+02].

Each approach are high throughput and can provide precise, gene expression levels. However, significant portion of the short tags can not be uniquely mapped to the reference genome [WGS09].

RNA-Seq RNA sequencing, has clear advantages over existing approaches it uses deepsequencing technologies where a population of RNA is converted to a library of cDNA fragments with adaptors attached to one or both ends. Each molecule is then sequenced in a high-throughput manner to obtain short sequences from one or both ends [WGS09].

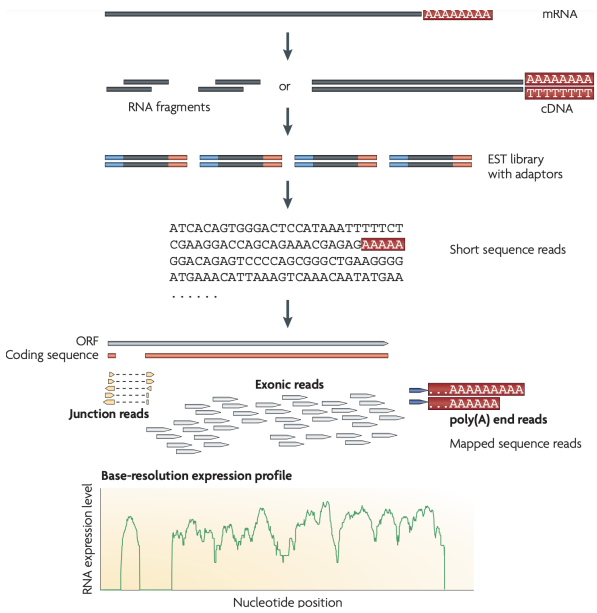


Figure 6: A typical RnA-seq experiment. RNAs are first converted into a library of cDNA fragments through either RNA fragmentation or DNA fragmentation. Sequencing adaptors are subsequently added to each cDNA fragment and a short sequence is obtained from each cDNA using high-throughput sequencing technology. The resulting sequence reads are aligned with the reference genome or transcriptome.

4 Alpha Fold

AlphaFolds goal is to predict the 3D coordinates of all heavy atoms for a given protein using the primary amino acid sequence and aligned sequences of homologues as inputs [JEP+21].

Mutations in proteins can lead to misfolding which is often associated with disease states, for example, Alzheimer's and Parkinson's which is one of the chanllanges for alphaFold [Fel].

The output is a file containing the 3D coordinates for every non-hydrogen atom in the protein. whilst showing the confidence levels for every amino acid residue, providing the reliability of the predicted structure [Fel].

Bioinformatics with Alpha Fold

Look for how alpha fold started Find more about how alpha fold works In July 2021, AlphaFold was developed by DeepMind, and was made available to the public [TAW+21].

Where it tries to silve the issue of invariant protein structures that are under translations and rotations [BP03].

AlphaFold is trained on protein chains from the PDB using the input sequence to query databases of protein sequences to generate a multiple sequence alignment [JEP+21]. Although we still do no exactly know how a protein sequence fold and alphafold does not help in figureing this out but its impact will likely be in accelerating and improving production of new medications [NZLJ22].

4.1 AlphaFold 2

The CASP14 was recently held which is a blind trial that critically assesses techniques for protein structure prediction [DITS22], AlphaFold2 was entered and out-performed all comptetitors.

Recently, RoseTTAFold was developed, trying to implement similar principles. Since then, other end-to-end structure predictors have emerged using different principles such as fast multiple sequence alignment processing in DMPFold218 and language model representations.[BPE22].

We use the root mean square deviation, to calculate the similarity between two structures, AlphaFold models had a accuracy of 0.96 compared to 2.80 which was the second best score. AlphaFold

models also had a high level of accuracy in predicting the position of residue side chains when the protein backbone prediction was accurate [DITS22] [JEP+21].

5 Implication for Bioinformatics

Bioinformatics concerns itself with the analysis of protein sequence to predict the structures, as well as its relationship to other proteins resulting in predicting the functionality. [Zve08].

Bibliography

[ALFJ+17] Rebecca F. Alford, Andrew Leaver-Fay, Jeliazko R. Jeliazkov, Matthew J. O'Meara, Frank P. DiMaio, Hahnbeom Park, Maxim V. Shapovalov, P. Douglas Renfrew, Vikram K. Mulligan, Kalli Kappel, Jason W. Labonte, Michael S. Pacella, Richard Bonneau, Philip Bradley, Roland L. Jr. Dunbrack, Rhiju Das, David Baker, Brian Kuhlman, Tanja Kortemme, and Jeffrey J. Gray. The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *Journal of Chemical Theory and Computation*, 13(6):3031–3048, June 2017. Publisher: American Chemical Society.

[BL22] Sarah E. Biehn and Steffen Lindert. Protein Structure Prediction with Mass Spectrometry Data. *Annual Review of Physical Chemistry*, 73(1):1–19, 2022. _eprint: https://doi.org/10.1146/annurev-physchem-082720-123928.

[Bou18] Boundless. 2.10: Atoms, Isotopes, Ions, and Molecules - Hydrogen Bonding and Van der Waals Forces, July 2018.

[BP03] Pierre Baldi and Gianluca Pollastri. The Principled Design of Large-Scale Recursive Neural Network Architectures–DAG-RNNs and the Protein Structure Prediction Problem. *NaN*, page 28, 2003.

[BPE22] Patrick Bryant, Gabriele Pozzati, and Arne Elofsson. Improved prediction of protein-protein interactions using AlphaFold2. *Nature Communications*, 13:1265, March 2022.

[BT98] Carl Ivar Branden and John Tooze. *Introduction to Protein Structure*. Garland Science, New York, 2 edition, December 1998.

[DBB03] Cyril Dominguez, Rolf Boelens, and Alexandre M. J. J. Bonvin. HADDOCK: A Protein Protein Docking Approach Based on Biochemical or Biophysical Information. *Journal of the American Chemical Society*, 125(7):1731–1737, February 2003. Publisher: American Chemical Society.

[DITS22] Alessia David, Suhail Islam, Evgeny Tankhilevich, and Michael J. E. Sternberg. The AlphaFold Database of Protein Structures: A Biologist's Guide. *Journal of Molecular Biology*, 434(2):167336, January 2022.

[EWMR+06] Narayanan Eswar, Ben Webb, Marc A. Marti-Renom, M. S. Madhusudhan, David Eramian, Min-Yi Shen, Ursula Pieper, and Andrej Sali. Comparative protein structure modeling using Modeller. *Current Protocols in Bioinformatics*, Chapter 5:Unit–5.6, October 2006.

[Fel] Felix. A brief introduction to AlphaFold | Science | Felix Online.

[God22] W T. Godbey. Chapter 3 - Proteins. In W T. Godbey, editor, *Biotechnology and its Applications (Second Edition)*, pages 47–72. Academic Press, January 2022.

[JEP+21] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger,

Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, August 2021. Number: 7873 Publisher: Nature Publishing Group.

[KKN+06] Rimantas Kodzius, Miki Kojima, Hiromi Nishiyori, Mari Nakamura, Shiro Fukuda, Michihira Tagami, Daisuke Sasaki, Kengo Imamura, Chikatoshi Kai, Matthias Harbers, Yoshihide Hayashizaki, and Piero Carninci. CAGE: cap analysis of gene expression. *Nature Methods*, 3(3):211–222, March 2006. Number: 3 Publisher: Nature Publishing Group.

[KMY+15] Lawrence A. Kelley, Stefans Mezulis, Christopher M. Yates, Mark N. Wass, and Michael J. E. Sternberg. The Phyre2 web portal for protein modeling, prediction and analysis. *Nature Protocols*, 10(6):845–858, June 2015. Number: 6 Publisher: Nature Publishing Group.

[LLV18] Patanachai Limpikirati, Tianying Liu, and Richard W. Vachet. Covalent labeling-mass spectrometry with non-specific reagents for studying protein structure and interactions. *Methods (San Diego, Calif.)*, 144:79–93, July 2018.

[LWL+20] Julia Koehler Leman, Brian D. Weitzner, Steven M. Lewis, Jared Adolf-Bryfogle, Nawsad Alam, Rebecca F. Alford, Melanie Aprahamian, David Baker, Kyle A. Barlow, Patrick Barth, Benjamin Basanta, Brian J. Bender, Kristin Blacklock, Jaume Bonet, Scott E. Boyken, Phil Bradley, Chris Bystroff, Patrick Conway, Seth Cooper, Bruno E. Correia, Brian Coventry, Rhiju Das, René M. De Jong, Frank DiMaio, Lorna Dsilva, Roland Dunbrack, Alexander S. Ford, Brandon Frenz, Darwin Y. Fu, Caleb Geniesse, Lukasz Goldschmidt, Ragul Gowthaman, Jeffrey J. Gray, Dominik Gront, Sharon Guffy, Scott Horowitz, Po-Ssu Huang, Thomas Huber, Tim M. Jacobs, Jeliazko R. Jeliazkov, David K. Johnson, Kalli Kappel, John Karanicolas, Hamed Khakzad, Karen R. Khar, Sagar D. Khare, Firas Khatib, Alisa Khramushin, Indigo C. King, Robert Kleffner, Brian Koepnick, Tanja Kortemme, Georg Kuenze, Brian Kuhlman, Daisuke Kuroda, Jason W. Labonte, Jason K. Lai, Gideon Lapidoth, Andrew Leaver-Fay, Steffen Lindert, Thomas Linsky, Nir London, Joseph H. Lubin, Sergey Lyskov, Jack Maguire, Lars Malmström, Enrique Marcos, Orly Marcu, Nicholas A. Marze, Jens Meiler, Rocco Moretti, Vikram Khipple Mulligan, Santrupti Nerli, Christoffer Norn, Shane Ó'Conchúir, Noah Ollikainen, Sergey Ovchinnikov, Michael S. Pacella, Xingjie Pan, Hahnbeom Park, Ryan E. Pavlovicz, Manasi Pethe, Brian G. Pierce, Kala Bharath Pilla, Barak Raveh, P. Douglas Renfrew, Shourya S. Roy Burman, Aliza Rubenstein, Marion F. Sauer, Andreas Scheck, William Schief, Ora Schueler-Furman, Yuval Sedan, Alexander M. Sevy, Nikolaos G. Sgourakis, Lei Shi, Justin B. Siegel, Daniel-Adriano Silva, Shannon Smith, Yifan Song, Amelie Stein, Maria Szegedy, Frank D. Teets, Summer B. Thyme, Ray Yu-Ruei Wang, Andrew Watkins, Lior Zimmerman, and Richard Bonneau. Macromolecular modeling and design in Rosetta: recent methods and frameworks. *Nature Methods*, 17(7):665–680, July 2020. Number: 7 Publisher: Nature Publishing Group.

[LZG20] Xiaoran Roger Liu, Mengru Mira Zhang, and Michael L. Gross. Mass Spectrometry-Based Protein Footprinting for Higher-Order Structure Analysis: Fundamentals and Applications. *Chemical Reviews*, 120(10):4355–4454, May 2020. Publisher: American Chemical Society.

[NZLJ22] Ruth Nussinov, Mingzhen Zhang, Yonglan Liu, and Hyunbum Jang. AlphaFold, Artificial Intelligence (AI), and Allostery. *The Journal of Physical Chemistry B*, 126(34):6372–6383, September 2022. Publisher: American Chemical Society.

[OM06] Michał J. Okoniewski and Crispin J. Miller. Hybridization interactions between probesets in short oligo microarrays lead to spurious correlations. *BMC Bioinformatics*, 7(1):276, June 2006.

[OR15] Robert J. Ouellette and J. David Rawn. 14 - Amino Acids, Peptides, and Proteins. In Robert J. Ouellette and J. David Rawn, editors, *Principles of Organic Chemistry*, pages 371–396. Elsevier, Boston, January 2015.

[oV] University of Vermont. Levels of Protein Organization.

[RBL+02] Jeannette Reinartz, Eddy Bruyns, Jing-Zhong Lin, Tim Burcham, Sydney Brenner, Ben Bowen, Michael Kramer, and Rick Woychik. Massively parallel signature sequencing (MPSS) as a tool for in-depth quantitative gene expression profiling in all organisms. *Briefings in Functional Genomics*, 1(1):95–104, February 2002.

[RLW+12] Daniel Russel, Keren Lasker, Ben Webb, Javier Velázquez-Muriel, Elina Tjioe, Dina Schneidman-Duhovny, Bret Peterson, and Andrej Sali. Putting the Pieces Together: Integrative Modeling Platform Software for Structure Determination of Macromolecular Assemblies. *PLOS Biology*, 10(1):e1001244, January 2012. Publisher: Public Library of Science.

[RRG07] Thomas E. Royce, Joel S. Rozowsky, and Mark B. Gerstein. Toward a universal microarray: prediction of gene expression through nearest-neighbor probe sequence identification. *Nucleic Acids Research*, 35(15):e99, 2007.

[SC10] Harry W Schroeder and Lisa Cavacini. Structure and Function of Immunoglobulins. *The Journal of allergy and clinical immunology*, 125(2 0 2):S41–S52, February 2010.

[SL20] Justin T. Seffernick and Steffen Lindert. Hybrid methods for combined experimental and computational determination of protein structure. *The Journal of Chemical Physics*, 153(24):240901, December 2020. Publisher: American Institute of Physics.

[TAW+21] Kathryn Tunyasuvunakool, Jonas Adler, Zachary Wu, Tim Green, Michal Zielinski, Augustin Žídek, Alex Bridgland, Andrew Cowie, Clemens Meyer, Agata Laydon, Sameer Velankar, Gerard J. Kleywegt, Alex Bateman, Richard Evans, Alexander Pritzel, Michael Figurnov, Olaf Ronneberger, Russ Bates, Simon A. A. Kohl, Anna Potapenko, Andrew J. Ballard, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Ellen Clancy, David Reiman, Stig Petersen, Andrew W. Senior, Koray Kavukcuoglu, Ewan Birney, Pushmeet Kohli, John Jumper, and Demis Hassabis. Highly accurate protein structure prediction for the human proteome. *Nature*, 596(7873):590–596, August 2021. Number: 7873 Publisher: Nature Publishing Group.

[WGS09] Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63, January 2009. Number: 1 Publisher: Nature Publishing Group.

[YYR+15] Jianyi Yang, Renxiang Yan, Ambrish Roy, Dong Xu, Jonathan Poisson, and Yang Zhang. The I-TASSER Suite: protein structure and function prediction. *Nature Methods*, 12(1):7–8, January 2015. Number: 1 Publisher: Nature Publishing Group.

[Zve08] Marketa J. Zvelebil. *Understanding bioinformatics / Marketa Zvelebil & Jeremy O. Baum.* Garland Science/Taylor & Francis Group, Garland Science, Taylor & Francis Group, New York, 2008.