Using grammars to analyze the relation between protein folding and sequence in low complexity regions

Andrade M., Biology
Schmid F., Soft Matter Physics

While large parts of proteins fold into structured domains, which are fairly conserved in evolution, and become relatively easy to isolate and study, it is becoming more and more evident that an important part of protein function resides in low complexity regions (LCRs) (Mier et al., 2019).

These LCRs are very abundant but are difficult to study for their structure and function. They evolve very quickly and they are less conserved in evolution than globular domains. In addition, they seem to adopt disordered structures and the techniques that are used to experimentally characterize protein structure do not work with them. Methods that predict intrinsically disordered regions tend to identify LCRs as such.

However, there is increasing evidence that LCRs might provide flexible regions that adopt structure upon interactions with protein partners. Our own work on homorepeats of glutamines (polyQ) demonstrate that these form a gradient of temporary helical structure that decreases from N to C- terminal. Various reports suggest that they extend an N-terminal coiled coil upon interaction with another coiled-coil from a partner protein (e.g. Petrakis et al. 2013).

Moreover, we identify many LCRs with high repeatability (sequences with short perfect repeats or just some mutations away from them) in all eukaryotic proteomes examined. Differences observed between species suggest particular distributions of short repeats associated to some species (Kamel et al., 2019).

Together, this evidence suggests that many LCRs might gain function by approaching repeatability, which could result in structure. However, predicting such structure, moreover if this is flexible and dependent on context of interaction and surrounding structure, is an unsolved problem.

Here, we propose to explore possibilities to use the computer science concept of grammars to identify rules that govern the construction of sequences and establish correlations between sequences and folding. Simply speaking, grammars in computer science are replacement rules that are used to generate sequences of objects, which may or may not depend on the context of the variables that are being replaced.

Our project will consist of two complementary parts.

In the first part (main supervisor M. Andrade), the student will search for grammar rules that may underlie amino acid sequences in LCRs. The underlying idea is that the evolution of LCRs may proceed by certain rules (e.g., facilitated insertion of certain repeats at certain positions of proteins, facilitated removal of certain other sequences), which constrain the resulting sequence space. Our preliminary work already gives indications that this is the case for glutamine rich regions (Urbanek et al., under revision).

In the second part (main supervisor F. Schmid), we will use a toy protein model to explore possible relations between sequence and folding structures. To this end, we will use lattice heteropolymer models, which have been popular models for generic studies of the statistical physics of protein folding for many decades (see Bogner et al and Behringer et al for own related work). In these models, proteins are represented by chains of a very reduced set of "amino acids", e.g., H (hydrophobic), "P" (polar), and "O" (other), which are confined to a lattice. Therefore, an exhaustive analysis of the statistical properties of the corresponding ensemble becomes possible. For example,

in the protein folding problem, it has been shown that there exist "highly designable" conformations that can be realized by a large number of sequences (Bialek).

Our idea in the project is to define certain basic grammar rules for our toy sequences that will be inspired by the structures of the repeats in real LCRs sequences, and investigate the relation between the grammar rules and the (statistically averaged) structural properties of the resulting heteropolymers. Variations in the grammar will be tested.

This simplified setup should help to provide folding rules for sequences with very short repeats depending on the underlying grammar (which determines the relations between repeats). Ideally, we will optimize grammar rules for small protein repeats that provide the maximum amount of ordered structure, based on the assumption that evolution has been searching for repeats with defined structures.

This project should answer the questions of why particular amino acid combinations are found in small repeats in particular species, and most importantly, will provide dynamic structural predictions for LCRs, reproducing the alpha-helical tendency in polyQ.

References:

Kamel M, Mier P, Tari A, **Andrade-Navarro MA**. Repeatability in protein sequences. J Struct Biol. 2019 Aug 10. pii: S1047-8477(19)30173-X. doi: 10.1016/j.jsb.2019.08.003.

Mier P, Paladin L, Tamana S, Petrosian S, Hajdu-Soltész B, Urbanek A, Gruca A, Plewczynski D, Grynberg M, Bernadó P, Gáspári Z, Ouzounis CA, Promponas VJ, Kajava AV, Hancock JM, Tosatto SCE, Dosztanyi Z, **Andrade-Navarro MA**. Disentangling the complexity of low complexity proteins. Brief Bioinform. 2019 Jan 30. doi: 10.1093/bib/bbz007.

Petrakis S, Schaefer MH, Wanker EE, **Andrade-Navarro MA**. Aggregation of polyQ-extended proteins is promoted by interaction with their natural coiled-coil partners. Bioessays. 2013 Jun;35(6):503-507. doi: 10.1002/bies.201300001. Epub 2013 Mar 11.

Urbanek A, Popovic M, Morató A, Estaña A, Elena-Real CA, Mier P, Fournet A, Allemand F, Delbecq S, **Andrade-Navarro MA**, Cortés J, Sibille N, Bernadó P.

2019. Flanking regions determine the structure of the poly-glutamine homo-repeat in huntingtin through a mechanism that is common in human proteins. In revision in Nature Struct. Mol. Biol.

Behringer, H, Bogner, T, Polotsky, A, Degenhard, A, and **Schmid, F**, Developing and analyzing idealized models for molecular recognition, J. Biotechnology 129, 268 (2006).

Bogner, T, Polotsky, A, Degenhard, A, and **Schmid, F,** Molecular recognition in a lattice model: An enumeration study**.** Phys. Rev. Lett. 93, 268108 (2005).

William Bialek, Biophysics: Searching for principles, chapter 5.1, Princeton University Press (2012).