Modelling the EGFR through the Protein Conformation Ontology*

Giacomo De Colle* [1,3], Morgan Mitchell[2], Alexander D. Diehl[2,3]

[1] *Department of Philosophy, University at Buffalo, Buffalo (NY), US*
[2] *Department of Biomedical Informatics, University at Buffalo, Buffalo (NY), US*
[3] *National Center for Ontological Research*

Abstract

The study of the spatial configuration of the structures of a protein, also known as protein conformation, is pivotal to the understanding of the functioning and activation patterns of proteins, as well as their relations with the surrounding environment. In order to facilitate data-driven research on protein conformations, we present the Protein Conformation Ontology, an ongoing project which provides a structured vocabulary of terms used to represent protein conformations and related conformational changes at different levels of granularity. To the aim of testing the capabilities of the ontology, we adopted as a test case the conformational changes related to the activation of the epidermal growth factor receptor. In this paper, we discuss our initial results in modelling two different models of the epidermal growth factor receptor.

Keywords

Work-in-progress, Protein Conformation Ontology (PRC), biomedical ontologies, protein conformations, Basic Formal Ontology (BFO))

1. Introduction

The spatial configuration of a protein is a key element to our understanding of its functioning. The same protein, depending on the different environmental conditions it is situated in, can fold into many different types of structures, also known as protein conformations [1]. Differences in folding patterns and conformational changes of a protein can bring it to be involved in radically different types of processes. Protein conformations can be classified at different levels of granularity called the secondary, tertiary, and quaternary levels. At the secondary level, protein structures identify hydrogen bonding between atoms in the same polypeptide backbone. Tertiary structures are more complex

SeWebMeDA-2024: 7th International Workshop on Semantic Web Solutions for Large-scale Biomedical Data Analytics, May 26, 2024, Hersonissos, Greece
* Corresponding author.
gdecolle@buffalo.edu (G. De Colle); addiehl@buffalo.edu (A. D. Diehl); mm528@buffalo.edu (M. Mitchell)
0000-0002-3600-6506 (G. De Colle); 0000-0001-9990-8331 (A. D. Diehl); 0000-0002-9318-128X (M. Mitchell)
© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

CEUR-WS.org/Vol-3726/paper4.pdf

and are identified with the way in which one polypeptide chain disposes itself in three-dimensional space when side chains interact with one another. Quaternary structures are created when two or more polypeptide chains connect [2].

Ontologies are controlled vocabularies developed for allowing consistent semantic interoperability of large and fragmented datasets [3]. This paper presents developments in creating the Protein Conformation Ontology (PRC), the first attempt at building an ontology that is able to represent all types of protein conformations, their changes over time and their relation to protein functioning. The PRC can be used to build representations of secondary, tertiary, and quaternary structures, thus enabling querying and comparison of protein databases on the basis of protein conformation. We present the current state of development of the PRC and focuses on modelling the epidermal growth factor receptor (EGFR) as a use case for the ontology.

A survey of BioPortal [4] and of the ontologies included in the OBO Foundry [5] reveal that existing ontologies representing protein structures or conformations include the Protein Ontology (PRO), the Sequence Ontology (SO), the Semanticscience Integrated Ontology (SIO) and the Physico-Chemical Methods and Properties ontology (FIX). Nevertheless, all these efforts only represent certain aspects of protein structure and cannot be used as comprehensive tools to model conformational changes. PRO, for example, cannot be used to represent secondary and tertiary protein structure, although it does represent proteoforms, i.e. protein isoforms and post-translational modifications that represent variants in the structure of a given protein [6].

SO on the other hand represents protein structure at the secondary structure level, but cannot be used to model protein conformations at an higher level of complexity. Moreover, the scope of SO is limited to continuous sequences of amino acids, thus effectively disregarding representation of secondary structures that include discontinuous regions [7]. Some common protein secondary conformations are also included in SIO and FIX, but both ontologies have a relatively narrow scope and lack many terms, especially in the realm of tertiary and quaternary protein conformation [8].

The scope of the ontologies described above is then too narrow to fully represent the full variety of protein conformations, as well as changes in protein conformations, and warrants the creation of an ontology devoted to this particular domain. In this paper we present first results in developing the PRC, an ontology which intends to address this issue. The aim of the PRC is not only to represent protein conformations as physical structure or as structured material entities. Rather, the PRC represents conformational changes, the way in which these conformational changes take place, the way in which they are triggered and the way in which they are ordered in time. For this reason, the PRC identifies protein conformations as dispositions to adopt a certain structure. Such dispositions are realized in a process of conformational change, where the protein adopts the corresponding to the protein conformation.

2. Methods

The development of the PRC originated from the need to represent the conformational changes involved in the formation of protein aggregates in neurological diseases, such as the formation of prions associated with Creutzfeldt-Jakob disease [9]. PRC was built using the Stanford Center for Biomedical Informatics Research tool, Protégé (version 5.6.1) [10]. The HermiT 1.4.3.456 [11] reasoner was used to check the ontology for logical consistency. The PRC adopts the Basic Formal Ontology (BFO) as a top-level architecture. Moreover, the PRC imports terms from the Gene Ontology (GO), the Protein Ontology (PRO) and the Relation Ontology (RO) [3 6 12 13]. SO served as a valuable source of information about protein secondary structures, but SO in ambiguous as to whether its classes represent information content entities or material entities, and as such is not compliant with the BFO. Many PRC terms are based on SO terms, but have definitions rewritten to emphasize that the conformations represented are types of dispositions. Reference to the original terms in SO was given by using the annotation properties 'definition source' from the Information Artifact Ontology (IAO) and skos:closematch. It should be noted that PRC includes many more secondary structure classes than SO because of our more thorough curation and representation of secondary structures formed from discontinuous sequence regions, and we will propose matching classes in SO for those that are conformant with SO's representation approach (which allows for only continuous sequences of amino acids).

PRC is developed with the aim of supporting and complying with FAIR data standards practices [14]. When possible, resources imported by other ontologies have been used as described in the paragraph above in order to allow for reuse of data, and they have been referenced by providing metadata by using, for example, skos:closematch. The aim of the project is to develop the ontology in compliance with OBO Foundry principles [15], and to submit a future version of the ontology for acceptance to the OBO community, thus allowing for better findability of the ontology. At the moment, the latest version of the PRC ontology can be found at the following GitHub page:

https://github.com/Buffalo-Ontology-Group/Protein-Conformation-Ontology

The development process of the PRC began with the identification of a class of entities of interest for research in the biomedical domain, i.e. protein conformations, and the consideration of use cases where the ontological representation of these entities would have been useful. BFO was adopted as a top-level ontology from the beginning of the development process, and the PRC team decided to use the BFO class "disposition" to represent how conformations are created by processes which change the material structure of the proteins they inhere in. In this way, protein conformations can be natively connected to the functioning of proteins which they are responsible for. During the development process, we created models of use cases selected from scientific literature, which we then evaluated and used as a guide to introduce or modify the initial top-level structure we developed out of BFO. This paper focuses on one of such use-cases. A more complete presentation of the development of the PRC will be submitted elsewhere and include other use cases that we have explored, such as representing the conformational changes that the

voltage-gated sodium channel [16] is involved in and the conformational changes involved in the formation of protein aggregates implicated causally in various neurological diseases.

3. Results

The PRC currently contains 206 PRC-prefixed classes, which include terms representing secondary, tertiary, and quaternary conformation dispositions, as well as terms representing conformational changes, structural qualities, and material entities such as protein complexes and amino acid chains. Currently, the PRC includes many terms representing secondary structures, such as 'beta helix', that refer to conformations commonly found in many different protein complexes. Tertiary and quaternary structures, on the other hand, are very specific to particular protein complexes, and cannot all be represented in a single ontology. To test the capabilities of the ontology to represent tertiary and quaternary structures, we chose to represent certain specific protein complexes and their conformational changes over time and in different conditions.

The use-case we present in this paper as a test case for the PRC is the epidermal growth factor receptor (EGFR). The EGFR is a transmembrane protein responsible for regulation of a series of events such as coordination of cell growth, differentiation, and migration and which serves a role in epithelial development [17]. Other members of the EGFR protein superfamily, which share structural similarities with the EGFR, have roles in other parts of the body, such as cardiac development [17 18]. Errors in the activation of the EGFR are closely related to the formation of tumors [19]. Representing EGFR is a formidable test case for the PRC, given that the EGFR is involved in a very complex series of conformational changes including dimerization. Building an ontological model of the EGFR is then an excellent test case for any ontology that aims at representing conformations and conformational changes over time.

The EGFR is a tyrosine kinase receptor (RTK) which activates a process of phosphorylation that regulates cell production. The mechanism through which this process is activated is of particular interest for a structural study of proteins. The EGFR is activated through binding with one of its ligands, usually the epidermal growth factor (EGF), and through a complex process of conformational changes, that in some cases involve ligand-induced dimerization with another EGFR monomer [17 18 19].

Two models based on a variety of experimental data and simulations have been developed to represent the mechanism through which the EGFR activates. In one of the models, called "ligand-induced dimerization model", two EGFR subunits dimerize after binding to EGF. In the other model, called "rotation model", EGFR exists as a preformed dimer, and is merely activated after ligand binding occurs. Despite the rotation model having received stronger experimental confirmation in recent years, consensus has not been reached on whether the model should entirely replace the ligand-induced one [20 21]. We thus decided to include the ligand-induced and rotation models in our representation,

not only because scientific consensus has not yet been reached, but also because both models may reflect reality, in that some EGFR subunits may be pre-dimerized and some not [22].

The EGFR comprises an external region that is divided into four domains, includes a transmembrane domain and a juxtamembrane domain, that connects the external domains to the tyrosine kinase domain (TKD) to a disordered carboxyl tail [18]. Each of these parts is described by a 'material entity' class in the PRC, along with the secondary structures it is composed of.

The process through which the EGFR activates its TKD, according to the ligand-induced dimerization model, is the following: first of all, one EGF monomer binds to an EGFR monomer, between the external domains 1 and 3 [17]. This causes domain 3 to rotate of roughly 120 degrees closer to domain 1, and the whole external part of the EGFR to undergo a radical conformational change. A beta hairpin called "dimerization arm", part of external domain 2, is now pointing outwards and can be used to bind to another EGFR that has also undergone the same sequence of changes. The two dimerization arms must pass close to each other and bind reciprocally to a site on the external domain 2 of the opposite EGFR monomer, resulting in the formation of the external EGFR dimer [17].

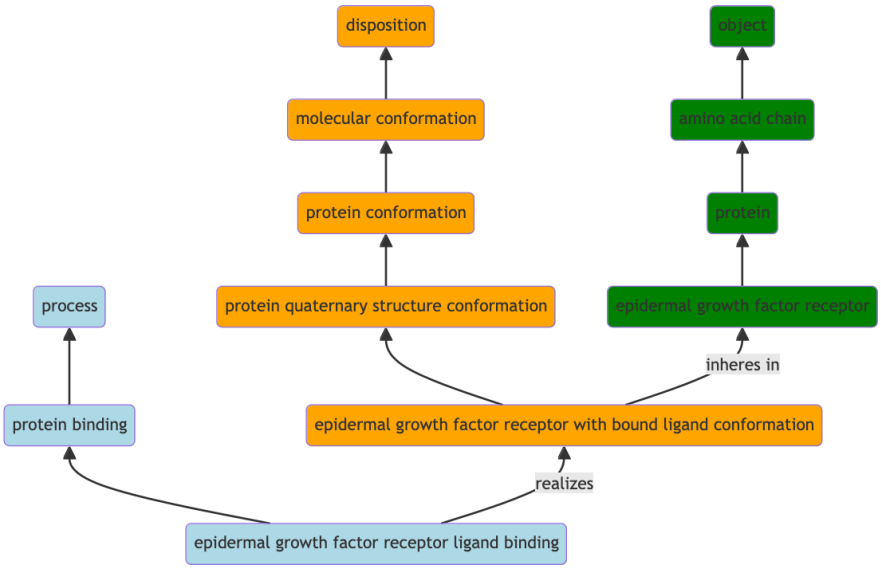

Figure 1: intended design pattern for PRC classes. Blue represents processes, green represents objects, orange represents dispositions. Processes realize dispositions, which inhere in objects.

As a consequence, other parts of the two EGFR monomers also bind. In particular, the intracellular TKD subunits dimerize and create an asymmetric TKD dimer. This dimer is formed through the rotation of an alpha helix in one TKD, called the alpha C helix, that interacts with an active site in the opposite TKD. As a result, the two TKDs, one assuming the role of an activator and the other the role of a receiver, will trans-phosphorylate each other.

These two processes of dimerization have been the main focus of our ontological representation. They represent a complex series of conformational changes that two different protein monomers undergo together in order to build a common quaternary structure. All of the entities described above have been represented in the ontology in the form of material entities for the parts of the EGFR, dispositions for the conformations of the EGFR, roles and processes of conformational changes. Processes in particular have been also axiomatized in order to automate reasoning, especially regarding their temporal ordering.

For example, the class “epidermal growth factor receptor ligand binding” has been made equivalent to

“'protein binding'
and (*realizes* **some**
'epidermal growth factor receptor with bound ligand conformation')
and (*precedes* **some**
'epidermal growth factor receptor activation')”.

When needed, new ad-hoc classes were created to reflect this different process ordering, as well as corresponding new disposition classes and relative axioms. When possible, classes from the ligand-induced dimerization model were reused, for example to represent the material entities involved in the constitution of the EGFR. Similarly, two different classes are introduced to represent the ligand-induced and rotation-induced processes of activation, and each have different processes as parts.

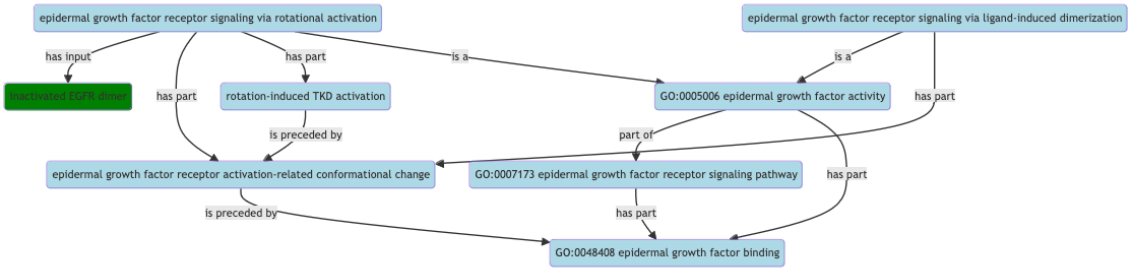

Figure 2: Process classes used for representing the two models of the EGFR in the PRC and their relationship with GO terms.

4. Discussion

The EGFR test case was successfully completed. The complex conformational changes that the protein complex is involved in are all represented in the PRC and properly axiomatized. Moreover, two different models for the EGFR functioning were also successfully represented and can now be coherently compared and used for querying purposes. Many subclasses of tertiary and quaternary structure were created to represent the structure of the EGFR protein complex: its disposition to connect with an EGFR ligand, its disposition to form a dimer with another EGFR monomer, its disposition to activate by rotating its dimerization arm, and so on. Further work in representing the EGFR will include linking the conformational changes in the ontology with other biological processes such as phosphorylation and extending our ontological representation of the normal forms of EGFR to include how pathological mutations alter its conformation and activation. Changes in the conformational activations of the EGFR might be related with its mis-activation in cases where the EGFR is involved with the formation of cancer [18]. Representing this process ontologically would provide an extremely helpful application of the PRC.

The PRC is aimed at representing protein conformations, how they are created as a result of various biological processes and how they inhere in different material entities. Allowing for querying of information about protein structures can also be beneficial a series of data-driven areas of research. For example, systems biology investigates biological functions based on the interactions between different parts of the whole organism. The ELIXIR platform has recently become a focal point for data-driven investigation in the field of systems biology [22]. Conformational changes depend on features of the biological environment that their bearers are situated in, and this is interestingly also the case for the EGFR [23]. Moreover, being able to differentiate between the two possible EGFR models on the basis of surrounding environmental features would reveal an extremely useful use case for an EGFR ontological representation.

5. Conclusion

The PRC is an ontology which adopts BFO as a top-level architecture and which aims at representing protein conformation, the processes they are involved in, and the material entities they depend on. The current development of the PRC includes 206 classes which represent secondary, tertiary, and quaternary structures. To test the capabilities of the PRC in representing protein conformations, we have modeled the EGFR, its dimerization process, and its activation, includes representing the two competing models of the EGFR in a coherent way. Being able to represent the complex series of conformational changes that interest the EGFR at different levels of granularity, the PRC proves to be the first successful ontology for modeling protein conformations and their changes.

Acknowledgements

The authors wish to acknowledge the helpful discussions with Lauren Wishnie, Darren Natale, and William Duncan.

References

[1] Anfinsen CB. The formation and stabilization of protein structure. Biochemical Journal **128**(4) (1972) 737-49. doi: 10.1042/bj1280737.

[2] Rehman I, Farooq M, Botelho S. "Biochemistry, Secondary Protein Structure". StatPearls. Treasure Island (FL), 2022.

[3] Arp R, Smith B, Spear AD. "Building Ontologies with Basic Formal Ontology", The MIT Press, 2015.

[4] Whetzel PL, Noy NF, Shah NH, et al. "BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications". Nucleic Acids Res 39 (2011).

[5] Jackson R, Matentzoglu N, Overton JA, et al. "OBO Foundry in 2021: operationalizing open data principles to evaluate ontologies". Database (Oxford) (2021). doi: 10.1093/database/baab069.

[6] Natale DA, Arighi CN, Blake JA, et al. "Protein Ontology (PRO): enhancing and scaling up the representation of protein entities". Nucleic Acids Res 45 (2017). doi: 10.1093/nar/gkw1075.

[7] Eilbeck K, Lewis SE, Mungall CJ, et al. "The Sequence Ontology: a tool for the unification of genome annotations. Genome Biology" 6(5):R44 (2005). doi: 10.1186/gb-2005-6-5-r44.

[8] Dumontier M, Baker CJ, Baran J, et al. "The Semanticscience Integrated Ontology (SIO) for biomedical research and knowledge discovery". J Biomed Semantics **5**(1) (2014). doi: 10.1186/2041-1480-5-14.

[9] Liu, F., Lü, W., & Liu, L. (2024). “New implications for prion diseases therapy and prophylaxis”. Frontiers in molecular neuroscience, 17, 1324702. https://doi.org/10.3389/fnmol.2024.1324702

[10] Noy NF, Crubezy M, Fergerson RW, et al. "Protégé-2000: an open-source ontology-development and knowledge-acquisition environment". AMIA Annu Symp Proc (2003).

[11] Glimm B, Horrocks I, Motik B, Stoilos G, Wang Z. "HermiT: An OWL 2 Reasoner". Journal of Automated Reasoning 53(3) (2014) 245-69. doi: 10.1007/s10817-014-9305-1.

[12] Consortium TGO. "The Gene Ontology Resource: 20 years and still GOing strong". Nucleic Acids Res 47(D1):D330-d38 (2014). doi: 10.1093/nar/gky1055.

[13] Smith B, Ceusters W, Klagges B, et al. "Relations in biomedical ontologies". Genome Biol **6**(5):R46 (2005). doi: 10.1186/gb-2005-6-5-r46.

[14] *Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J., Bonino da Silva Santos, L., Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S.C., Evelo, C.T., Finkers, R., González-Beltrán, A.N., Gray, A.J., Groth, P., Goble, C.A., Grethe, J.S., Heringa, J., 't Hoen, P.A., Hooft,*

R.W., Kuhn, T., Kok, R.G., Kok, J.N., Lusher, S.J., Martone, M.E., Mons, A., Packer, A.L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R.C., Sansone, S., Schultes, E.A., Sengstag, T., Slater, T., Strawn, G.O., Swertz, M.A., Thompson, M., van der Lei, J., van Mulligen, E.M., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., & Mons, B. (2016). "The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3".

[15] *Smith, B., Ashburner, M., Rosse, C., Bard, J.B., Bug, W.J., Ceusters, W., Goldberg, L.J., Eilbeck, K., Ireland, A., Mungall, C.J., Leontis, N.B., Rocca-Serra, P., Ruttenberg, A., Sansone, S., Scheuermann, R.H., Shah, N.H., Whetzel, P.L., & Lewis, S.E. (2007). "The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration". Nature Biotechnology, 25, 1251-1255.*

[16] Catterall WA. "Voltage-gated sodium channels at 60: structure, function and pathophysiology". The Journal of Physiology 590(11) (2012) 2577-89. doi: 10.1113/jphysiol.2011.224204.

[17] Ferguson KM. "Structure-based view of epidermal growth factor receptor regulation". Annu Rev Biophys. 37 (2008) 353-73. doi: 10.1146/annurev.biophys.37.032807.125829. PMID: 18573086; PMCID: PMC2745238.

[18] Lemmon MA, Schlessinger J, Ferguson KM. "The EGFR family: not so prototypical receptor tyrosine kinases". Cold Spring Harb Perspect Biol. 2014 Apr 1;6(4):a020768. doi: 10.1101/cshperspect.a020768. PMID: 24691965; PMCID: PMC3970421.

[19] Zhang X, Gureasko J, Shen K, Cole PA, Kuriyan J. "An allosteric mechanism for activation of the kinase domain of epidermal growth factor receptor". Cell. 125(6) (2006) 1137-49. doi: 10.1016/j.cell.2006.05.013. PMID: 16777603.

[20] Purba ER, Saita EI, Maruyama IN. "Activation of the EGF Receptor by Ligand Binding and Oncogenic Mutations: The "Rotation Model"". Cells. 6(2) (2017) 13 doi: 10.3390/cells6020013. PMID: 28574446; PMCID: PMC5492017.

[21] Zanetti-Domingues LC, Korovesis D, Needham SR, Tynan CJ, Sagawa S, Roberts SK, Kuzmanic A, Ortiz-Zapater E, Jain P, Roovers RC, Lajevardipour A, van Bergen En Henegouwen PMP, Santis G, Clayton AHA, Clarke DT, Gervasio FL, Shan Y, Shaw DE, Rolfe DJ, Parker PJ, Martin-Fernandez ML. "The architecture of EGFR's basal complexes reveals autoinhibition mechanisms in dimers and oligomers". Nat Commun. 9(1):4325. (2018) doi: 10.1038/s41467-018-06632-0. PMID: 30337523; PMCID: PMC6193980.

[22] Arkhipov A, Shan Y, Das R, Endres NF, Eastwood MP, Wemmer DE, Kuriyan J, Shaw DE. "Architecture and membrane interactions of the EGF receptor". Cell 31;152(3) (2013) 557-69. doi: 10.1016/j.cell.2012.12.030. PMID: 23374350; PMCID: PMC3680629

[23] Martins Dos Santos V, Anton M, Szomolay B, et al. "Systems Biology in ELIXIR: modelling in the spotlight". F1000Research 2022;**11**:1265 doi: 10.12688/f1000research.126734.1.

[24] Lelimousin M, Limongelli V, Sansom MS. "Conformational Changes in the Epidermal Growth Factor Receptor: Role of the Transmembrane Domain Investigated by Coarse-Grained MetaDynamics Free Energy Calculations". J Am Chem Soc. 138(33) (2016) 10611-22. doi: 10.1021/jacs.6b05602. Epub 2016 Aug 11. PMID: 27459426; PMCID: PMC5010359.