4. Our recent solutions

There are several motivations for solving exactly the computational protein design systems for which a reasonably good solution can be found with confidence. Thereby, quickly as problems get harder. Additionally, the mean energy gap to optimality tends to the optimal solution. The routine may end up trapped in local minima far from the global one. To avoid this problem, stochastic optimization is used. However, the current interface design methods used by the Rosetta software created in 2012 by the teams of David Baker and Todd Yeates. Later on, the same teams pushed the design approach even further and created larger assemblies of two-component systems (Bale et al. 2016). When dealing with large assemblies, both, computationally and evolutionarily, it is preferable to work with high-order symmetries. Progress in using proteins as the building blocks has been much more challenging, owing in part to the much greater complexity of the rules that govern the native structures of proteins. Nonetheless, nature has achieved spectacular assemblies using protein molecules as building blocks. Various examples range from viral capsids to microtubules and molecular carriers.

2. The need for multi-component systems

In order to design a novel protein to self-assemble into a complex but well-defined architecture, the protein molecule must contain multiple self-associating interfaces (King 2012). It turns out, somewhat surprisingly that two distinct self-associating interfaces is sufficient to create a wide range of outcomes, from cages to extended three-dimensional materials. A successful computational approach was demonstrated in 2012 by the teams of David Baker and Todd Yeates. Later on, the same teams pushed the design approach even further and created larger assemblies of two-component systems (Bale 2016). Indeed, larger sizes of designed assemblies are only possible if multiple non-identical protein components are present in the asymmetric unit.

3. The limitations of traditional approaches

Undoubtedly, the current design pipeline is very expensive both computationally and experimentally. For example, currently, there are no efficient ways to pre-scan protein interfaces of known folds that would satisfy the space-group constraints imposed by the desired design (private communications with Todd Yeates in April 2019 at the CAPRI protein docking conference). This fact significantly reduces the choice of protein folds to be used in interface design.

Also, the current interface design methods used by the Rosetta software created in the Baker’s team use stochastic optimization techniques and all-atom potentials. Although they have the advantage of providing the best known solution at any time, they neither guarantee finding the global minimum of the energy surface (i.e. GMEC: Global Minimum-Energy Conformation) in finite time nor a bounded energetic distance to the optimal solution. The routine may end up trapped in local minima far from the global one. To avoid this problem, stochastic optimization is used. However, the accuracy of stochastic methods drastically degrades as problem size increases (Voigt et al. 2000; Simoncini et al. 2015) and the probability of finding the GMEC drops very quickly as problems get harder. Additionally, the mean energy gap to optimality tends to increase with the number of designed residues, putting a limit on the size of systems for which a reasonably good solution can be found with confidence. Thereby, there are several motivations for solving exactly the computational protein design (CPD) problem.

2020-02711 - PhD Position F/M Algorithms for computational protein design

Type de contrat : CDI
Niveau de diplôme exigé : Bac + 5 ou équivalent
Fonction : Docteurant

A propos du centre ou de la direction fonctionnelle

Grenoble Rhône-Alpes Research Center groups together a few less than 800 people in 38 research teams and 8 research support departments.

Staff is localized on 5 campuses in Grenoble and Lyon, in close collaboration with labs, research and higher education institutions in Grenoble and Lyon, but also with the economic players in these areas.

Present in the fields of software, high-performance computing, Internet of things, image and data, but also simulation in oceanography and biology, it participates at the best level of international scientific achievements and collaborations in both Europe and the rest of the world.

Contexte et atouts du poste

Within the framework of a partnership

- collaboration between Nano-D team of Inria and INRA Toulouse

Mission confiée

For a better knowledge of the proposed research subject :

1. The need for proteins and high-order symmetries

Considerable progress has been demonstrated in the last 30 years in designing DNA (and RNA) sequences to form structures and materials (see, e.g., DNA origami). The solution turned out to be rather computationally efficient. However, biochemically, it is very challenging to work with DNAs and RNAs and it is much more preferable to create designs based on proteins. At the same time, an outstanding goal in bioengineering is to design macromolecules that assemble into complex higher-order structures (Bale et al. 2016). When dealing with large assemblies, both, computationally and evolutionarily, it is preferable to work with high-order symmetries.

Progress in using proteins as the building blocks has been much more challenging, owing in part to the much greater complexity of the rules that govern the native structures of proteins. Nonetheless, nature has achieved spectacular assemblies using protein molecules as building blocks. Various examples range from viral capsids to microtubules and molecular carriers.

2. The need for multi-component systems

In order to design a novel protein to self-assemble into a complex but well-defined architecture, the protein molecule must contain multiple self-associating interfaces (King 2012). It turns out, somewhat surprisingly that two distinct self-associating interfaces is sufficient to create a wide range of outcomes, from cages to extended three-dimensional materials. A successful computational approach was demonstrated in 2012 by the teams of David Baker and Todd Yeates. Later on, the same teams pushed the design approach even further and created larger assemblies of two-component systems (Bale et al. 2016). Indeed, larger sizes of designed assemblies are only possible if multiple non-identical protein components are present in the asymmetric unit.

3. The limitations of traditional approaches

Undoubtedly, the current design pipeline is very expensive both computationally and experimentally. For example, currently, there are no efficient ways to pre-scan protein interfaces of known folds that would satisfy the space-group constraints imposed by the desired design (private communications with Todd Yeates in April 2019 at the CAPRI protein docking conference). This fact significantly reduces the choice of protein folds to be used in interface design.

Also, the current interface design methods used by the Rosetta software created in the Baker’s team use stochastic optimization techniques and all-atom potentials. Although they have the advantage of providing the best known solution at any time, they neither guarantee finding the global minimum of the energy surface (i.e. GMEC: Global Minimum-Energy Conformation) in finite time nor a bounded energetic distance to the optimal solution. The routine may end up trapped in local minima far from the global one. To avoid this problem, stochastic optimization is used. However, the accuracy of stochastic methods drastically degrades as problem size increases (Voigt et al. 2000; Simoncini et al. 2015) and the probability of finding the GMEC drops very quickly as problems get harder. Additionally, the mean energy gap to optimality tends to increase with the number of designed residues, putting a limit on the size of systems for which a reasonably good solution can be found with confidence. Thereby, there are several motivations for solving exactly the computational protein design (CPD) problem.

Informations générales

- Thème/Domaine : Biologie numérique
- Ville : Grenoble
- Centre Inria : CRI Grenoble - Rhône-Alpes
- Date de prise de fonction souhaitée : 2020-11-01
- Durée de contrat : 3 ans
- Date limite pour postuler : 2020-09-30

Contacts

- Equipe Inria : NANO-D
- Directeur de thèse : Grudinin Sergei / sergei.grudinin@inria.fr

A propos d’Inria

Inria est l’institut national de recherche dédié aux sciences et technologies du numérique. Il emploie 2600 personnes. Ses 200 équipes-projets agiles, en général communes avec des partenaires académiques, impliquent plus de 3500 scientifiques pour relever les défis du numérique, souvent à l’interface d’autres disciplines. L’institut fait appel à de nombreux talents dans plus d’une quarantaine de métiers différents. 900 personnes d’appui à la recherche et à l’innovation contribuent à faire émerger et grandir des projets scientifiques ou entrepreneuriaux qui impactent le monde. Inria travaille avec de nombreuses entreprises et a accompagné la création de plus de 180 start-up. L’institut s’efforce ainsi de répondre aux enjeux de la transformation numérique de la science, de la société et de l’économie.

L’essentiel pour réussir

We are looking for a candidate who is passionate about applying mathematical models to biological objects on the atomic scale.

Consignes pour postuler

Sécurité défense :

Ce poste est susceptible d’être affecté dans une zone à régime restrictif (ZRR), telle que définie dans le décret n°2011-1425 relatif à la protection du potentiel scientifique et technique de la nation (PPST). L’autorisation d’accès à une zone est délivrée par le chef d’établissement, après avis ministériel favorable, tel que défini dans l’arrêté du 03 juillet 2012, relatif à la PPST. Un avis ministériel défavorable pour un poste affecté dans une ZRR aurait pour conséquence l’annulation du recrutement.

Politique de recrutement :

Dans le cadre de sa politique diversité, tous les postes Inria sont accessibles aux personnes en situation de handicap.

Attention : Les candidatures doivent être déposées en ligne sur le site Inria. Le traitement des candidatures adressées par d’autres canaux n’est pas garanti.
Several exact deterministic approaches guaranteeing that, if run to completion, the returned solution is the GMEC have been proposed. They mainly rely on the Dead-End Elimination (DEE) theorem (Desnét et al., 1992), the A* algorithm (Leach et al., 1998), other branch and bound techniques (Gordon et al., 2003; Hong et al., 2009), integer linear programming (Kingsford et al., 2005) and dynamic programming (Leaver-Fay et al., 2004). Guaranteed deterministic methods are the only methods which offer a provable basis for improving biophysical models. Indeed, they ensure that discrepancies between CPD predictions and experimental results come exclusively from modelling inadequacies and not from the algorithm. These properties are crucial to rationally tune the energy-based scoring functions (Alvizo and Mayo, 2008). Unfortunately, these methods are often rapidly outstripped by the complexity of the search space and do not provide any solution.

Thanks to INRA partner’s (LISBP & MIAT) recent work, this has now changed. We have tackled these algorithmic challenges by adapting, extending and experimenting new algorithms proposed in artificial intelligence to the specific combinatorial optimization problems inherent to CPD. Our developments have led to new computational biological design approaches based on graphical models and more specifically Cost Function Network technology (CFN aka Weighted CSP; implemented in the toolbar2 solver, Cooper et al., 2010) that enable efficient handling of complex sequence-conformation spaces previously unsolvable by state-of-the-art provable CPD methods (Aliouche et al., 2012; Traoré et al., 2013; Aliouche et al., 2014; Simoncini et al., 2015, Traoré et al., 2016; Traoré et al., 2017; Viricel et al. 2018). These CFN-based methods rely on Local Consistency filtering, a family of CFN pruning and incremental lower bounding techniques (Cooper et al., 2010) combined with Branch and Bound enhanced with variable elimination and graph-based problem decomposition techniques. Impressively, compared to classical methods, the CFN-based approaches speed-up by several orders of magnitude the search process and provide a guaranteed GMEC for much larger CPD problems than were previously attainable. Ultimately, CPD problems that could not be solved using hundreds of CPU-hours on computer clusters using traditional provable methods can now be solved to optimality in a few minutes on a laptop using CFN-based approaches. In addition to finding the proven optimal solution, these latter also enable the exhaustive enumeration of ensembles of near-optimal solutions that are often unattainable using other methods. These impressive progresses provide new routes to the exact solving of CPD optimization problems with runtime performances that compete against those of heuristics while guaranteeing optimality. A recent achievement using these highly effective methods was the design of a highly stable artificial self-assembling protein (symmetrical eight-bladed β-propeller of 319 amino acids) whose structure has been validated by x-ray crystallography and different biophysical methods (Noguchi et al. 2019).

Thesis conditions:
The PhD student will be supervised by Dr Sergei Grudinin (Inria / CNRS) and Dr Sophie Barbe (INRA-LISBP). Regarding model and methods developments, the project will also benefit from expertise and know-how of Dr T. Schiex (INRA-MIAT), and Dr. J. Esque (INRA-LISBP). Regarding experimental evaluation and validation, C. Montanier (INRA-LISBP) at LISBP will be involved.

The project will benefit from molecular modeling software/computing equipment and scientific environment available at LISBP-INRA Toulouse and Nano-D Inria Grenoble as well as computing resources and support provided by TGIR such as the GENCI High Performance Computing facilities TGIR resources, the Computing Mesocenter of the Region Midi-Pyrénées (CALMIP, Toulouse), and the GenoToul Bioinformatics Platform of INRA-Toulouse. Experimental facilities at LISBP, will be used for the experimental validation of protein design methods.

Principales activités
The overall goal of the current PhD proposal is to push protein design method even further and apply it to symmetric multi-component systems by combining expertise of the partners involved:

- Computational design method developments (the first half of the thesis, main contributor - Inria).
  - The first goal of the proposal will be to extend the rapid fast-Fourier transform-accelerated method developed by the Inria partner for assemblies with space-group symmetries.
  - The second goal of the proposal will be to study the conformational variability of individual protein subunits under various symmetry and crystal packing constraints.
  - The third goal of the proposal is to develop coarse-grained potential for protein design and to collaboratively integrate it into CPD methods developed by the INRA partners.
- Computational design method evaluation and validation (the second half of the thesis, main contributor - INRA).
  - The ultimate goal of the proposal will be to apply the developed methods to practical designs of single-and multi-component systems.

Compétences

Technical skills and level required:
We require strong knowledge of applied math, linear algebra, computer science and machine learning. Knowledge and understanding of statistical physics is a plus. The candidate will work with C++, python and models in structural biology.

Languages: The working language is English, French is a plus.

Avantages

- Subsidized meals
- Partial reimbursement of public transport costs
• Leave: 7 weeks of annual leave + 10 extra days off due to RTT (statutory reduction in working hours) + possibility of exceptional leave (sick children, moving home, etc.)
• Possibility of teleworking (after 6 months of employment) and flexible organization of working hours
• Professional equipment available (videoconferencing, loan of computer equipment, etc.)
• Social, cultural and sports events and activities
• Access to vocational training
• Social security coverage

Rémunération
Salary (before taxes) : 1982€ gross/month for 1st and 2nd year. 2085€ gross/month for 3rd year.