J Comput Aided Mol Des (2017) 31:287–291
DOI 10.1007/s10822-016-9988-y

Enabling drug discovery project decisions with integrated computational chemistry and informatics

Vickie Tsui[1] · Daniel F. Ortwine[1] · Jeffrey M. Blaney[1]

Received: 29 July 2016 / Accepted: 22 October 2016 / Published online: 31 October 2016

Abstract Computational chemistry/informatics scientists and software engineers in Genentech Small Molecule Drug Discovery collaborate with experimental scientists in a therapeutic project-centric environment. Our mission is to enable and improve pre-clinical drug discovery design and decisions. Our goal is to deliver timely data, analysis, and modeling to our therapeutic project teams using best-in-class software tools. We describe our strategy, the organization of our group, and our approaches to reach this goal. We conclude with a summary of the interdisciplinary skills required for computational scientists and recommendations for their training.

Keywords Computational chemistry · Drug design · Genentech · Cheminformatics · Software · Drug · Discovery

Introduction

The Computational Drug Discovery (CDD) group is in the Department of Discovery Chemistry, along with medicinal, purification, and analytical chemistry groups. Our group handles the curation and storage of all assay and chemistry data as well as building and maintaining the applications used to search and analyze these data. We are responsible for preclinical informatics, from collecting and registering experimental data, to experimental request systems, and to disseminating these data to therapeutic project team scientists. We also analyze these data and perform modeling with a combination of commercial, open-source, and in-house software. Because Genentech small molecule discovery goals require chemists and other scientists to be able to personally perform data analysis as well as structure- and property-based design, our group is focused on deploying informatics and computational tools to the desktops of all discovery scientists.

Therapeutic project teams include representatives from all pre-clinical small molecule drug discovery groups. The project team is responsible for defining the team's goals, developing the team's plan, and executing the plan. Our computational chemists are integral therapeutic project team members, responsible for helping drive scientific strategy via data analyses and modeling.

This perspective starts with the philosophy and strategy for how our group operates. We then describe software tools deployed to enable chemists and other scientists to analyze structure–activity relationships (SAR) and design small molecules. We discuss the make-up of the CDD group, and how that fits into our project-centric environment by maximizing the impact of tools to address the needs of therapeutic project teams. To conclude, we provide a perspective on the education necessary for computational chemistry graduate students to succeed in today's pharmaceutical/biotechnology company environment.

Philosophy and strategy

Computational scientists in drug discovery research are employed to maximize the impact of lab scientists by minimizing the number of compounds and experiments needed to advance compounds into clinical trials. Our therapeutic project team focus makes our role clear: to

✉ Jeffrey M. Blaney
blaney.jeff@gene.com

[1] Discovery Chemistry, Genentech, Inc., 1 DNA Way,
South San Francisco, CA 94080, USA

collaborate and help drive the teams' strategy, tactics, and decision-making. Drug discovery is a highly iterative process: analyze data, develop a hypothesis for what must be improved in the molecules to achieve the desired therapeutic compound profile, design molecules to test the hypothesis, synthesize the molecules, test the molecules, and repeat. Computational chemistry, modeling, informatics, and automated workflows play key roles in all of these iterative steps. Fundamentally, the CDD group's main job is to help project teams decide which experiments to do next to get to a good molecule faster.

Our main strategy to achieve this is to combine computational chemistry, cheminformatics, and software engineering into a single group. Embedding "software–savvy" Small Molecule Drug Discovery (SMDD) computational chemistry and informatics scientists into therapeutic project teams allows them to benefit from (or suffer with) the same software tools and infrastructure as other team scientists. In this way, they can rapidly identify problems, propose solutions, and work with other CDD group members to implement them, to the benefit of all project teams. The CDD group is located physically and organizationally within Discovery Chemistry to ensure a direct connection between medicinal chemists, other project team scientists, and CDD group members.

Sophisticated computational chemistry and modeling is useless if our teams cannot access all the key data required for analysis, design, and decision-making. We therefore prioritize informatics first and modeling second. Basic informatics includes database design and development, compound registration, data collection, data transfer, data transformation and loading, assay registration, workflow tools (e.g. assay requests), electronic lab notebook (ELN) support and integration, and searching and analysis tools. Cheminformatics includes virtual library enumeration and analysis, similarity, substructure, and SAR analysis. Computational chemistry includes physicochemical property estimations, small molecule conformational and electronic structure analysis, protein homology modeling, ligand and structure-based design. The skills and software tools required to perform basic informatics, cheminformatics, and computational chemistry overlap a great deal. Our approach integrates all of these activities into a single group, where most of the group is embedded into specific therapeutic project teams.

CDD group members' performance metrics are very similar to medicinal chemists': What specific impact did they achieve for their project that made a critical difference to the team? For example, our baseline expectation for a computational chemist is that their work must lead to specific experiments. Supporting therapeutic project team scientists by merely performing tasks they request is insufficient. Our group members' performance reviews and promotions are primarily driven by feedback from their therapeutic project team members, not their line management within the CDD group.

Much of SMDD experimental work is performed outside of Genentech by collaborators in different continents. We work closely with computational chemists at these external groups. Genentech SMDD frequently chooses to share project team research data with our collaborators. The CDD group therefore focuses on developing an agile, modular software infrastructure that not only supports our internal project teams, but also a frequently changing mix of external collaborators.

Our group supports most aspects of preclinical small molecule research informatics. We provide tools for compound and assay registration, assay data capture, and facilitate storage of these data in our corporate database. We extract and transform these data for delivery into project team-specific analysis and modeling tools. We maintain assay request (developed in-house) and fee-for-service contract research organization (CRO) synthesis request [1] systems, as well as commercial reagent and screening compound searching, purchase, and inventory [2, 3] software. We provide ELNs to internal [4] and external [5] chemists. We also conduct patent analyses and assist the legal department in assembling patents.

We deploy commercial and in-house software to all SMDD scientists, and provide focused training on how to actually do science (i.e., structure-based design) using those software. This involves close collaboration with therapeutic project team scientists and software companies. Training is conducted by CDD group members, project team experts, software company trainers, and external consultants. We prefer to purchase applications or use open source software whenever robust and flexible options are available. This provides agility to the group to react to new technologies and deliver "best-in-class" informatics and modeling approaches to our project teams. Because group members are largely freed from maintaining homegrown internal applications, their effort can be directed at software integration, exploring novel methodologies, and most importantly, to helping therapeutic project teams make better decisions. Software support of our infrastructure and integration tools still consumes a considerable fraction of our time, but this strategy helps minimize the support effort.

A critical part of our strategy involves choosing software vendors who work collaboratively with us. We work with these companies to develop and improve the software we license and allow them to incorporate these improvements in future software releases. This increases our capabilities over time and hopefully benefits the drug discovery field in general. We meet frequently with vendor scientists and programmers to provide direct feedback on their current products and future plans. Likewise, CDD members are encouraged to participate in vendor user group meetings, frequently as speakers.

Computational tools

Our group is also responsible for all computational modeling in SMDD. This includes conformational analysis, physicochemical property calculation, structure and ligand-based design, and quantitative structure–activity relationship (QSAR) and machine-learning QSAR and/or quantitative structure–property relationship (QSPR) model building and deployment. We integrate additional informatics and computational chemistry tools into Vortex [6], a chemistry-aware two-dimensional spreadsheet and plotting tool, and MOE [7] for three-dimensional molecular modeling and design. We linked Vortex to MOE to dynamically display 3D structures of molecules that have been highlighted in Vortex spreadsheets and plots. We deploy these software tools to internal and CRO scientists, and train them in project-specific contexts.

Our approach is to use commercial and open source software tools and toolkits that we identify as best in class, then build software to integrate them. We also develop new methods internally to supplement commercially available tools, including basic command line tools that can be used in Linux/Unix scripts or KNIME [8] workflows. We use a modular architecture to develop scalable, extensible, documented and supportable solutions. This modular architecture facilitates replacing software components when we identify new, superior approaches. We rely on robust open source software packages such as R [9], KNIME [8], and a wide variety of Java libraries and applications for property estimations and to integrate our tools. With the intention of stimulating the development of validated and reusable tools, we have released novel methods as Open Source [10] and regularly contribute to open source packages [11].

We have integrated many of the software tools into a few packages to minimize the number of user interfaces that our scientists must learn [12]. This requires licensing software that is highly modular with well-documented application programming interfaces (APIs) and web services. Such integration enables chemists to dock compounds, modify a ligand from a starting crystal structure, use the molecule as a query for interactive ligand-based screening for compounds that are available in-house or readily purchased, calculate properties on the resulting hits, and plot properties against each other in multiple dimensions. Additional capabilities include fast strain energy calculations and the computation of quantum mechanical torsion profiles [13], comparison of a selected torsion angle with the experimental profile from the Cambridge Structural Database (CSD) [14] and internal small molecule X-ray structures, and viewing various types of protein–ligand interactions, including dipolar and halogen interactions [15].

MOE is our primary three-dimensional desktop molecular modeling tool, used by CDD computational chemists, medicinal chemists, crystallographers, and other scientists. Ligand-based screening is enabled via a web-based interface built on top of FastROCS shape-feature searching [16]. Results are first previewed in a webpage containing two-dimensional Grapheme [17] depictions of the hit conformations' ability to reproduce the query features. Hits can then be imported directly into MOE, automatically superimposed onto the original query molecule or fragment(s).

Experimental data and calculated properties are analyzed in Vortex [6]. Biochemical, biological, DMPK (distribution, metabolism, pharmacokinetics), crystal structure, experimental and estimated property data are delivered to therapeutic project teams via custom-built Vortex session files that are auto-updated at least nightly. CDD group members create and customize these in collaboration with project team members to display relevant data, plots and graphs for that team. Similarly, MOE sessions are also generated and customized for each project by CDD group members, and include relevant and aligned crystal structures, models, and/or pharmacophores. The customization of these Vortex and MOE sessions can be time-consuming, which brings us to the next section that covers in detail the organization, roles and responsibilities of CDD group members.

CDD group organization

CDD scientists and software engineers work together to evaluate and purchase, build, and integrate best in class tools. Rather than dividing the CDD group into sub-teams (e.g. "application scientists", "cheminformaticians", "method developers", etc.), most members work across these functions. Approximately 25 % of our 19 group members focus primarily on collaborating with therapeutic project teams as a computational chemist, in addition to interacting with external software vendors and supporting one or more software applications. Another 35 % split their time 50/50 as computational chemists on project teams and as software engineers working on infrastructure, integration, and development. CDD software engineers are responsible for SMDD database development, improvement, maintenance, and integration with other applications, such as our chemistry ELN. They also support other SMDD groups, for example DMPK, analytical, and purification, with request systems and other tools. They work closely with other CDD group members and our central information technology (IT) group to support the collaborator data-sharing infrastructure, including taking responsibility for data transfer between Genentech and collaborating companies.

Each computational chemist works on two therapeutic projects at a time, or in the case of an individual whose time is split 50/50 between computational chemistry project support and software engineering, just one therapeutic project. We keep this ratio at ~0.5 computational chemist per project team throughout the various stages of a project, from inception through development candidate nomination and back-up compound research. Each computational chemist needs to be entrenched in their therapeutic team project(s) as a full collaborator, responsible for understanding all project aspects. They need to know and master the SAR, target biology, patent literature, structural biology, physicochemical properties, in vitro DMPK, and in vivo pharmacokinetics/pharmacodynamics (PK/PD) data. The computational chemist is involved in numerous processes from high-throughput screening (HTS) analysis, structure-based, ligand-based, and fragment-based design, to selection of compounds for PK and crystallography. The computational chemist is also responsible for ensuring the informatics support is appropriate and timely for their project team. Our collective group experience, supported by other Senior Discovery Chemistry leadership, is that working on more than two projects at a time reduces the impact that a computational chemist can have on team progress and risks turning their role into a service, rather than a proactive collaborator. Embedding the "50/50" CDD members with software engineering backgrounds and computational chemistry skills in project teams puts that person "in the trenches" with chemists, and frequently leads to development of novel computational tools or ways of integrating applications that are more likely to have direct project impact. The development and integration of these tools into MOE and Vortex during the last several years have enabled medicinal chemists, structural biologists, and other scientists to personally handle many of the more routine modeling and data analysis tasks formerly performed by computational chemists. This has freed up our computational chemists to focus on more specialized analyses and modeling, while also "raising the bar" for performance expectations.

Several members have joint appointments with other departments, including Early Discovery, DMPK, and Bioinformatics. This increases the interdisciplinary nature of the CDD group and helps ensure close collaborations are maintained. Three of our group members are heavily involved in building and refining models for physicochemical and DMPK properties [18]. Project teams use these models for compound design and prioritization. Another group member collaborates on early stage target validation, builds and expands screening libraries [19], evaluates drugability of new targets, and evaluates external lead-finding technologies.

CDD group members gain a wide breadth of knowledge in the drug discovery process by being involved from conception of a target to delivery of candidates for early development. We recruit and train computational chemists with the expectation that they will have broad expertise in all aspects of drug discovery, not just their core areas of computational chemistry and/or cheminformatics. A consequence of this is that several CDD members who aspire to be therapeutic Project Team Leaders (PTLs) have been given this opportunity. PTLs need to obtain deep understanding in all issues of the project, whether it is protein purification or off-target activities. These needs can lead to areas a computational chemist may not realize, such as discovery of "hidden data" that do not make it into the database and stay in team members' PowerPoint slides. In addition, in a monthly gathering of PTLs, new technologies are reviewed and different groups' capabilities are showcased, focusing on how they can impact project teams. CDD members who participate in these forums bring valuable information back to the rest of the CDD group, again leading to increased potential for direct in silico contributions to project progress.

Conclusions and perspectives

We described the goals, strategy, and approaches of Genentech's CDD group to enable all project team chemists to analyze their data and design small molecules themselves with software tools previously only available to computational chemists or "power users". A key part of the strategy is for "software–savvy" SMDD computational chemistry and informatics scientists to work directly as collaborators on therapeutic project teams.

Combining computational chemistry, cheminformatics, and software engineering in a single group makes us more responsive to the frequently changing and evolving needs of research. The interdisciplinary nature of the CDD group at Genentech is apparent from the descriptions provided above. While joint appointments with other departments and PTLs within the group demonstrate this, group members' overlapping expertise in computational chemistry, medicinal chemistry, and software engineering is a major reason for our positive recognition within the SMDD organization. Several group members are fully capable of supporting a project team, as well as developing robust software. These members can identify computational and informatics needs on their projects and, rather than having to go to someone else, design and write the code themselves or easily collaborate with other programmers in the group. Top-quality candidates with this combined computational chemist/software engineer phenotype are very difficult to find and recruit. Even though they may not be formally trained as software engineers, our computational chemists are highly software and informatics "savvy".

The demand for this type of computational chemistry professional, i.e., one with medicinal chemistry, drug design, and programming expertise, has been increasing. Many start-up companies need a single computational person to build up informatics and computational infrastructure and perform structure-based and ligand-based design at the same time. We therefore recommend that the training of graduate students in computational chemistry-related labs should encourage or even require students to gain experience in science *and* programming. Some professors have already been doing this, and more should consider this approach. Some labs often consist of students from a variety of departments, some of which do not have programming course requirements. A Ph.D. student's thesis project may involve only application of existing software without participating in the coding or even scripting of any software packages, or a solid understanding of the underlying algorithms and methods. Students frequently do not understand the practical application or implications of their computational studies: What experiments would they propose based on their results? Alternatively, a student can be trained in a computer science department, working only on software development for their thesis without taking science classes. Such scenarios can limit the student's scope of job search. On the other hand, more encouragement from computational chemistry professors for graduate students to tap into multiple disciplines, especially software engineering, physical organic chemistry, structural biology, and computational biology, will increase the pool of candidates who can straddle multiple tasks. These are the ones who may suggest a compound to synthesize one day, and code an improved scoring function the next. The demand for such multi-talented individuals will only grow in the future.

Acknowledgements We have summarized the ideas and contributions of our entire group in this perspective: Ignacio Aliagas, Paul Beroza, Fabio Broccatelli, Huifen Chen, Kevin P. Clark, Paul Gibbons, Alberto Gobbi, Chandra Goliva, Matthew Lardy, Man-Ling Lee, Chinchih Lu, Hans Purkey, Ben Sellers, Nick Skelton, Ryan Weekley, BinQing Wei, and Hao Zheng. We would also like to thank J.W. Feng for his software integration work and "desktop modeling" contributions. We thank Professor François Diederich and Professor Klaus Müller for their workshops and consulting. We benefited from joint discussions between our group and Roche computational chemists, especially Bernd Kuhn, Caterina Bissantz, Johannes Hermann, Rama Kondru, Harald Mauser, and Martin Stahl. Their initial customization of MOE led us to migrate our initial "desktop modeling" efforts from PyMOL [20] to the more modeling and design-oriented MOE. We are grateful to them for initiating the collaboration with Chemical Computing Group to streamline MOE for medicinal chemists. We thank Neil Taylor for his work to extend and customize Proasis' software to become a critical part of our infrastructure. We are grateful to our IT department for their support. Our group's ability to impact Medicinal Chemistry and Project Teams required clear goals and consistent support from Sr. Research Leadership, especially Bruce Roth, Wendy Young, Marcel Hop, Cris Lewis, and Mike Varney.

References

1. KSRS is available from Kelaroo. www.kelaroo.com
2. KRMS is available from Kelaroo. www.kelaroo.com
3. www.emolecules.com
4. Biovia Workbook is available from Biovia. www.accelrys.com
5. Studies Notebook is available from Dotmatics Ltd. www.dotmatics.com/products/vortex
6. Vortex is available from Dotmatics Ltd. www.dotmatics.com/products/vortex
7. Molecular Operating Environment (MOE) is available from Chemical Computing Group Inc. www.chemcomp.com
8. KNIME. www.knime.com
9. R Core Team (2015) R: a language and environment for statistical computing. R foundation for statistical computing, Vienna, Austria. http://www.R-project.org/
10. Gobbi A, Giannetti AM, Chen H, Lee ML (2015) Atom-atom-path similarity and sphere exclusion clustering: tools for prioritizing fragment hits. J Cheminform 7:11
11. https://sourceforge.net/projects/aestel/
12. Feng JA, Aliagas I, Bergeron P, Blaney JM, Bradley EK, Koehler MFT, Lee M-L, Ortwine DF, Tsui V, Wu J, Gobbi A (2015) An integrated suite of modeling tools that empower scientists in structure- and property-based drug design. J Comput Aided Mol Des 29:511–523
13. Lee M, Aliagas IM, Feng JA, Gabriel T, O'Donnell TJ, Sellers BD, Wiswedel B, Gobbi A (2016) Chemalot: a Command-Line Cheminformatics Open-Source Package, GitHub Repository. https://github.com/chemalot/chemalot
14. MOGUL is available from the Cambridge crystallographic data centre. www.ccdc.cam.ca.uk/products/csd_system/mogul/
15. Kuhn B, Fuchs JE, Reutlinger M, Stahl M, Taylor NR (2011) Rationalizing tight ligand binding through cooperative interaction networks. J Chem Inf Model 51:3180–3198
16. FastROCS is available from OpenEye Scientific Software. www.eyesopen.com/fastrocs
17. The Grapheme Toolkit is available from OpenEye Scientific Software. www.eyesopen.com/grapheme-tk
18. Ortwine DF, Aliagas I (2013) Physicochemical and DMPK in silico models: facilitating their use by medicinal chemists. Mol Pharm 10:1153–1161
19. Beresini MH, Liu Y, Dawes TD, Clark KR, Orren L, Schmidt S, Turincio R, Jones SW, Rodriguez RA, Thana P, Hascall D, Gross DP, Skelton NJ (2014) Small-molecule library subset screening as an aid for accelerating lead identification. J Biomol Screen 19:758–770
20. The PyMOL molecular graphics system is available from Schrodinger LLC. www.pymol.org