Bibliographie

[1] Ponder J. et Richards F. M. (**1987**) *Tertiary templates for proteins : Use of packing criteria in the enumeration of allowed sequences for different structural classes. J. Mol. Biol.*, volume 193, pages 775–792.
cité pages 3 et 60

[2] Su A. et Mayo S.L. (**1997**) *Coupling backbone flexibility and amino acid sequence selection in protein design. Protein Science*, volume 9, pages 1701–1707.
cité page 3

[3] Finkelstein A. et Ptitsyn O. (**1977**) *Theory of protein molecule self-organization. i. thermodynamic parameters of local secondary structures in the unfolded protein chain. Biopolymers*, volume 16, pages 469–495.
cité page 3

[4] Janin J., Wodak S., Levitt M. et Maigret B. (**1978**) *Conformation of amino acid sidechains in proteins. Journal of Molecular Biology*, volume 125, pages 357–386.
cité page 3

[5] McGregor M., Islam S. et Sternberg M. (**1987**) *Analysis of the relationship between sidechain conformation and secondary structure in globular proteins. J. Mol. Biol.*, volume 198, pages 295–310.
cité page 3

[6] Dunbrack R. et Karplus M. (**1993**) *Backbone-dependent rotamer library for proteins. application to side-chain prediction. J. Mol. Biol.*, volume 230, pages 543–574.
cité page 3

[7] Dahiyat B. et Mayo S. (**1997**) *De novo protein design : fully automated sequence selection. Science*, volume 278, pages 82–87.
cité page 3

[8] Desjarlais J. et Handel T. (**1999**) *Side-chain and backbone flexibility in protein core design. J. Mol. Biol.*, volume 290, pages 305–318.
cité pages 3 et 4

[9] Druart K., Bigot J., Audit E. et Simonson T. (**2016**) *A hybrid monte carlo scheme for multibackbone protein design. Journal of Chemical Theory and Computation*, volume 12, pages 6035–6048.
cité pages 4, 33 et 60

[10] Dantas G., Corrent C., Reichow S., Havranek J., Eletr Z., Isern N., Kuhlman B., Varani G., Merritt E. et Baker D. (**2007**) *High-resolution structural and thermodynamic analysis of extreme stabilization of human procarboxypeptidase by computational protein design. Journal of Molecular Biology*, volume 366, pages 1209–1221.
cité page 4

[11] Kuhlman B., Ireton G., Varani G., Stoddard B. et Baker D. (**2003**) *Design of a novel globular protein fold with atomic-level accuracy. Science*, volume 302, pages 1364– 1368.
cité pages 4 et 27

[12] Davis I., Arendall W., Richardson D. et Richardson J. (**2006**) *The backrub motion: how protein backbone shrugs when a sidechain dances. Structure*, volume 14, pages 265–274.
cité page 4

[13] Georgiev, I., Lilien, R. H., Donald et B. R. (**2008**) *The minimized dead-end elimination criterion and its application to protein redesign in a hybrid scoring and search algorithm for computing partition functions over molecular ensembles. J. Comput. Chem.*, volume 29, pages 1527–1542.
cité pages 4, 60 et 80

[14] Smith C. et Kortemme T. (**2008**) *Backrub-like backbone simulation recapitulates natural protein conformational variability and improves mutant side-chain prediction. Journal of Molecular Biology*, volume 380, pages 742–756.
cité page 4

[15] Dahiyat B. et Mayo S. (**1996**) *Protein design automation. Protein Science*, volume 5, pages 895–903.
cité pages 4 et 27

[16] Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M. Jr., Ferguson, D. M. Spellmeyer, D. C., Fox, T., Caldwell, J. W., , Kollman et P. A. (**1995**) *A second generation force field for the simulation of proteins, nucleic acids and organic molecules. J. Am. Chem. Soc.*, volume 117, pages 5179–5197.
cité pages 5 et 93

[17] Brooks, B., Brooks III, C. L., Mackerell Jr., A. D., Nilsson, L., Petrella, R. J., Roux, B., Won, Y., Archontis, G., Bartels, C., Boresch, S., Caflisch, A., Caves, L., Cui, Q., Dinner, A. R., Feig, M., Fischer, S., Gao, J., Hodoscek, M., Im, W., Kuczera, K., Lazaridis, T., Ma, J., Ovchinnikov, V., Paci, E., Pastor, R. W., Post, C. B., Pu, J. Z., Schaefer, M., Tidor, B., Venable, R. M., Woodcock, H. L., Wu, X., Yang, W., York, D. M., Karplus et M. (**2009**) *CHARMM: The biomolecular simulation program. J. Comp. Chem.*, volume 30, pages 1545–1614.
cité pages 5 et 66

[18] Jorgensen W. et Rives J. T. (**1988**) *The OPLS force field for proteins. energy minimizations for crystals of cyclic peptides and crambin. J. Am. Chem. Soc.*, volume 110, pages 1657–1666.
cité page 5

[19] Christen M., Hünenberger P., Bakowies D., Baron R., Bürgi R., Geerke D., Heinz T., Kastenholz M., Kräutler V., Oostenbrink C., Peter C., Trzesniak D. et van Gunsteren W. (**2005**) *The GROMOS software for biomolecular simulation : Gromos05. J. Comp. Chem.*, volume 16, pages 1719–1751.
cité page 5

[20] Zollars ES., Marshall SA., et Mayo SL (**2006**) *Simple electrostatic model improves designed protein sequences. Protein Science*, volume 15, pages 2014–2018.
cité page 7

[21] Pokala N. et Handel T. (**2005**) *Energy functions for protein design : Adjustement with protein-protein complex affinities, models for the unfolded state, and negative design of solubility and specificity. J. Mol. Biol.*, volume 347, pages 203–227.
cité page 7

[22] Korkut A. et Hendrickson W. (**2009**) *A force field for virtual atom molecular mechanics of proteins. Proc. Natl. Acad. Sci. USA*, volume 106, pages 15667–72.
cité page 7

[23] Wesson L. et Eisenberg D. (**1992**) *Atomic solvation parameters applied to molecular dynamics of proteins in solution. Protein Science*, volume 1, pages 227–235.
cité page 10

[24] Rocchia W, Sridharan S, Nicholls A, Alexov E, Chiabrera A et Honig B. (**2002**) *Rapid grid-based construction of the molecular surface and the use of induced surface charge to calculate reaction field energies: applications to the molecular systems and geometric objects. J Comput Chem.*, volume 23(1), pages 128–37.
cité page 11

[25] Baker NA, Sept D, Joseph S, Holst MJ et McCammon JA. (**2001**) *Electrostatics of nanosystems: application to microtubules and the ribosome. Proc Natl Acad Sci USA*, volume 98(10), pages 10037–41.
cité page 11

[26] Born M. (**1920**) *Volumen und hydratationswarme der ionen. Z. Phys.*, volume 1, pages 45–48.
cité page 11

[27] Still W., Tempczyk A., Hawley R. et Hendrickson T. (**1990**) *Semianalytical tratment of solvataion for molecular mechanics and dynamics. J Am Chem Soc*, volume 112, pages 6127–29.
cité page 11

[28] Polydorides S., Amara N., Aubard C., Plateau P., Simonson T. et Archontis G. (**2011**) *Computational protein design with a generalized born solvent model: application to asparaginyl-trna synthetase. Proteins*, volume 79, pages 3448–3468.
cité pages 13 et 81

[29] Simonson T, Gaillard T, Mignon D, Schmidt am Busch M, Lopes A, Amara N et et al (**2013**) *Computational protein design: the Proteus software and selected applications. J Comput Chem*, volume 34, pages 2472–2484.
cité pages 13, 27, 79, 117 et 118

[30] Gaillard T et Simonson T (**2014**) *Pairwise Decomposition of an MMGBSA Energy Function for Computational Protein Design. J Comput Chem*, volume 35, pages 1371–1387.
cité pages 13, 29, 33, 67 et 75

[31] Isard C. Perry B. (**2012**) *Theory and Practice in Replica-Exchange Molecular Dynamics Simulation. ProQuest, UMI Dissertation Publishing.*
cité page 14

[32] Desmet J., Mayer M. D., Hazes B. et Lasters I. (**1992**) *The dead-end elimination theorem and its use in protein side-chain positioning. Nature*, volume 356, pages 539–542.
cité page 15

[33] Goldstein R. (**1994**) *Ecient rotamer elimination applied to protein side-chains and related spin glasses. Biophysical Journal*, volume 66, pages 1335–1340.
cité page 16

[34] Leach A. et Lemon A. (**1998**) *Exploring the conformational space of protein side chains using dead-end elimination and the A* algorithm. Proteins: Structure, Function, and Genetics*, volume 33, pages 227–239.
cité page 16

[35] Pierce N.A., Spriet J.A., Desmet J. et Mayo S.L. (**2000**) *Conformational splitting: a more powerful criterion for dead-end elimination. J. Comp. Chem.*, volume 21, pages 999–1009.
cité page 16

[36] Lilien R., Stevens B., Anderson A. et Donald B. (**2005**) *A novel ensemble-based scoring and search algorithm for protein redesign and its application to modify the substrate specicity of the gramicidin synthetase a phenylalanine adenylation enzyme. Journal of Computational Biology*, volume 12, pages 740–761.
cité page 16

[37] Schiex T. (**2000**) *Arc consistency for soft constraints. Principles and Practice of Constraint Programming*, volume 1894, pages 411–424.
cité page 18

[38] Allouche D., André I., Barbe S., Davies J., Givry S., Katsirelos G., O'Sullivan B., Prestwich S., Schiex T. et Traoré S. (**2014**) *Computational protein design as an optimization problem. Artificial Intelligence*, volume 212, pages 59–79.
cité pages 18, 60, 61 et 65

[39] Traoré, S., Allouche, D., André, I., De Givry, S., Katsirelos, G, Schiex, T, Barbe et S. (**2013**) *A new framework for computational protein design through cost function network optimization. Bioinformatics*, volume 27 (19), pages 2129–2136.
cité pages 18, 60, 61 et 65

[40] Wernisch L., Hery S. et Wodak S. (**2000**) *Automatic protein design with all atom forceelds by exact and heuristic optimization. Journal of Molecular Biology*, volume 301, pages 713–736.
cité pages 19, 61, 65 et 79

[41] Goldberg D. E. et Holland J. H. (**1988**) *Genetic algorithms and machine learning. Machine learning*, volume 3(2), pages 95–99.
cité page 20

[42] Cercignani, C., Lampis et M. (**1981**) *On the H-theorem for polyatomic gases. J Stat Phys*, volume 26, pages 4.
cité page 22

[43] Swendsen R. H. et Wang J. S. (**1986**) *Replica Monte Carlo simulation of spin glasses. Physical Review Letters*, volume 57, pages 2607–2609.
cité page 24

[44] Gainza P., Roberts K., Georgiev I., Lilien R., Keedy D., Chen C., Reza F., Anderson A., Richardson D., Richardson J. et Donald B. (**2013**) *Osprey: Protein design with ensembles, exibility and provable algorithms. Methods in Enzymology*, volume 523, pages 87–107.
cité page 27

[45] Polydorides S., Michael E., Mignon D., Druart K., Archontis G. et Simonson T. (**2016**) *Proteus and the design of ligand binding sites. Methods in Molecular Biology: Computational Design of Ligand Binding Proteins*, volume 1414, pages 77–97.
cité pages 27 et 60

[46] A. T. Brünger (**1992**) X-PLOR *version 3.1, A System for X-ray crystallography and NMR*. Yale University Press, New Haven.
cité pages 28, 33, 51, 66 et 93

[47] Simonson T., Ye-Lehmann S., Palmai Z., Amara N., Wydau-Dematteis S., Bigan E., Druart K., Moch C. et Plateau P. (**2016**) *Redesigning the stereospecicity of tyrosyltrna synthetase. Proteins*, volume 84, pages 240–253.
cité page 29

[48] Villa V., Mignon D., Polydorides S. et Thomas Simonson T. (**2017**) *Comparing pairwise-additive and many-body Generalized Born models for acid/base calculations and protein design. J. Comp. Chem.*, volume 38(28), pages 2396–2410.
cité pages 29, 30 et 93

[49] Mignon D, Panel N, Chen X, Fuentes E et Simonson T (**2017**) *Computational design of the Tiam1 PDZ domain and its ligand binding. J Chem Theory Comput*, volume 13, pages 2271–89.
cité pages 29 et 104

[50] Street A.G. et Mayo S.L (**1998**) *Pairwise calculation of protein solvent-accessible surface areas. Folding and Design*, volume 3, pages 253–258.
cité pages 29 et 66

[51] Lopes A., Aleksandrov A., Bathelt C., Archontis G. et Simonson T. (**2007**) *Computational sidechain placement and protein mutagenesis with implicit solvent models. Proteins*, volume 67, pages 853–867.
cité pages 29, 66 et 75

[52] Hawkins G., Cramer C.J. et Truhlar D. (**1995**) *Pairwise solute screening of solute charges from a dielectric medium. Chem. Phys. Lett.*, volume 246, pages 122–129.
cité page 30

[53] Schaefer M. et Karplus M (**1996**) *A comprehensive analytical treatment of continuum electrostatics. J. Phys. Chem*, volume 100, pages 1578–1599.
cité page 30

[54] Archontis G et Simonson T (**2005**) *A residue-pairwise Generalized Born scheme suitable for protein design calculations. J Phys Chem B*, volume 109, pages 22667–22673.
cité page 31

[55] Tuffery P., Etchebest C., Hazout S. et Lavery R (**1991**) *A new approach to the rapid determination of protein side chain conformations. Journal of Biomolecular Structure Dynamics*, volume 8, pages 1267–1289.
cité pages 33 et 67

[56] Krivov GG, Shapalov MV et Dunbrack RL (**2009**) *Improved prediction of protein side-chain conformations with SCWRL4. Proteins*, volume 77, pages 778–795.
cité pages 33, 67 et 116

[57] Gaillard T, Panel N et Simonson T (**2016**) *Protein sidechain conformation predictions with an MMGBSA energy function. Proteins*, volume 84, pages 803–819.
cité pages 33, 67, 93 et 116

[58] Madera M, Vogel C, Kummerfeld SK, Chothia C et Gough J (**2004**) *The SUPERFAMILY database in 2004: additions and improvements. Nucl Acids Res*, volume 32, pages D235–D239.
cité page 39

[59] Andreeva A, Howorth D, Brenner SE, Hubbard JJ, Chothia C et Murzin AG (**2004**) *SCOP database in 2004: refinements integrate structure and sequence family data. Nucl Acids Res*, volume 32, pages D226–229.
cité pages 40, 68 et 100

[60] Hughey R. et Krogh A. (**1995**) *SAM: Sequence alignment and modeling software system.*
cité page 40

[61] http://hmmer.org .
cité page 40

[62] Punta M., Coggill P., Eberhardt R., Mistry J., Tate J., Boursnell C., Pang N., Forslund K., Ceric G., Clements J., Heger A., Holm L., Sonnhammer E., Eddy S., Bateman A. et Finn R. (**2012**) *The pfam protein families database. Nucleic Acids Research*, volume 40, pages 290–301.
cité page 40

[63] Finn, R.D., Bateman, A., Clements J., Coggill P., Eberhardt R.Y., Eddy S.R., Heger A., Hetherington K., Holm L., Mistry J., Sonnhammer E.L.L., Tate J. et Punta M. (**2014**) *The Pfam protein families database. Nucleic Acids Research*, volume Issue 42, pages D222–D230.
cité page 40

[64] Henikoff S. et Henikoff J. (**1992**) *Amino acid substitution matrices from protein blocks. PNAS*, volume 89, pages 10915–10919.
cité page 41

[65] Altschul S., Madden T., Schaffer A., Zhang J., Zhang Z., Miller W. et Lipman D. (**1997**) *Gapped balst and psi-blast : a new generation of protein database search programs. Nucleic Acids Res*, volume 25, pages 3389–3402.
cité page 41

[66] Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K et Madden T.L. (**2008**) *BLAST+: architecture and applications. BMC Bioinformatics*, volume 10, pages 421.
cité page 41

[67] Durbin R., Eddy S. R., Krogh A. et Mitchison G. (**2002**) *Biological sequence analysis.* Cambridge University Press, Cambridge, United Kingdom.
cité pages 41 et 68

[68] Launay G., Mendez R., Wodak S. J. et Simonson T. (**2007**) *Recognizing protein-protein interfaces with empirical potentials and reduced amino acid alphabets. BMC Bioinf.*, volume 8, pages 270–291.
cité pages 42 et 68

[69] Mignon D et Simonson T (**2016**) *Comparing three stochastic search algorithms for computational protein design: Monte Carlo, Replica Exchange Monte Carlo, and a multistart, steepestdescent heuristic. J Comput Chem*, volume 37, pages 1781–1793.
cité pages 59 et 93

[70] Dantas G, Kuhlman B, Callender D, Wong M et Baker D (**2003**) *A Large Test of Computational Protein Design: Folding and Stability of Nine Completely Redesigned Globular Proteins. J Mol Biol*, volume 332, pages 449–460.
cité pages 60, 79 et 103

[71] G. L. Butterfoss et B. Kuhlman (**2006**) *Computer-based design of novel protein structures. Ann. Rev. Biophys. Biomolec. Struct.*, volume 35, pages 49–65.
cité page 60

[72] Lippow SM et Tidor B (**2007**) *Progress in computational Protein Design. Curr Opin Biotech*, volume 18, pages 305–311.
cité page 60

[73] Saven et J. G. (**2011**) *Computational protein design: engineering molecular diversity, nonnatural enzymes, nonbiological cofactor complexes, and membrane proteins. Curr. Opin. Chem. Biol.*, volume 15, pages 452–457.
cité page 60

[74] Feldmeier K et Hoecker B (**2013**) *Computational protein design of ligand binding and catalysis. Curr Opin Chem Biol*, volume 17, pages 929–933.
cité page 60

[75] Tinberg CE, Khare SD, Dou J, Doyle L, Nelson JW, Schena A et et al (**2013**) *Computational design of ligand-binding proteins with high affinity and selectivity. Nature*, volume 501, pages 212–218.
cité page 60

[76] Pokala N et Handel TM (**2004**) *Energy functions for protein design I: Efficient and accurate continuum electrostatics and solvation. Prot Sci*, volume 13, pages 925–936.
cité page 60

[77] Samish I, MacDermaid CM, Perez-Aguilar JM et Saven JG (**2011**) *Theoretical and computational protein design. Ann Rev Phys Chem*, volume 62, pages 129—149.
cité page 60

[78] Li Z., Yang Y., Zhan J., Dai L. et Zhou Y. (**2013**) *Energy Functions in De Novo Protein Design: Current Challenges and Future Prospects. Ann Rev Biochem*, volume 42, pages 315–335.
cité page 60

[79] Aleksandrov A, Polydorides S, Archontis G et Simonson T (**2010**) *Predicting the Acid/Base Behavior of Proteins: A Constant-pH Monte Carlo Approach with Generalized Born Solvent. J Phys Chem B*, volume 114, pages 10634–10648.
cité pages 60 et 93

[80] Polydorides S et Simonson T (**2013**) *Monte Carlo simulations of proteins at constant pH with generalized Born solvent, flexible sidechains, and an effective dielectric boundary. J Comput Chem*, volume 34, pages 2742–2756.
cité page 60

[81] K. Kilambi et J. J. Gray (**2012**) *Rapid calculation of protein pK_a values using Rosetta. BPJ*, volume 103, pages 587–595.
cité page 60

[82] Looger L. et Hellinga H. (**2001**) *Generalized dead-end elimination algorithms make largescale protein side-chain structure prediction tractable: Implications for protein design and structural genomics. Journal of Molecular Biology*, volume 307, pages 429–445.
cité page 60

[83] D. B. Gordon et S. L. Mayo (**1999**) *Branch-and-Terminate: a combinatorial optimization algorithm for protein design. Structure*, volume 7, pages 1089–1098.
cité page 60

[84] E. J. Hong, S. M. Lippow, B. Tidor et T. Lozano-Perez (**2009**) *Rotamer optimization for protein design through MAP estimation and problem size reduction. JCC*, volume 30, pages 1923–1945.
cité pages 60 et 61

[85] Zou J. et Saven J. G. (**2003**) *Using self-consistent fields to bias Monte Carlo methods with applications to designing and sampling protein sequences. J. Chem. Phys. 11*, volume 8, pages 3843–3854.
cité page 60

[86] D. Frenkel et B. Smit (**1996**) *Understanding molecular simulation.* Academic Press, New York.
cité pages 60, 61 et 80

[87] C. Chipot et A. Pohorille (**2007**) *Free energy calculations: theory and applications in chemistry and biology.* Springer Verlag, N.Y.
cité page 60

[88] Schmidt am Busch M, Mignon D et Simonson T (**2009**) *Computational protein design as a tool for fold recognition. Proteins*, volume 77, pages 139–158.
cité pages 61, 67, 69, 77, 81, 99 et 118

[89] M. Schmidt am Busch, A. Sedano et T. Simonson (**2010**) *Computational protein design: validation and possible relevance as a tool for homology searching and fold recognition. PLoS One*, volume 5(5), pages e10410.
cité pages 61, 67, 69, 77 et 81

[90] J. R. Norris (**1988**) *Markov chains.* Cambridge University Press.
cité page 61

[91] Metropolis N., Rosenbluth A., Rosenbluth M., Teller A. et Teller E. (**1953**) *Equation of state calculations by fast computing machines. The Journal of Chemical Physics*, volume 21, pages 1087–1092.
cité pages 61, 62, 63 et 64

[92] D. Frenkel et B. Smit (**1996**) *Understanding molecular simulation.* Academic Press, New York.
cité pages 61, 62 et 63

[93] Sugita, Y., Okamoto et Y. (**1999**) *Replica-exchange molecular dynamics method for protein folding. Chem. Phys. Lett. 31*, volume 4, pages 141–151.
cité page 61

[94] Kofke DA (**2002**) *On the acceptance probability of replica-exchange Monte Carlo trials. J Chem Phys*, volume 117, pages 6911–6914.
cité pages 61, 64 et 65

[95] Earl D. et Deem M. W. (**2005**) *Parallel tempering: theory, applications, and new perspectives. Phys. Chem. Chem. Phys.*, volume 7, pages 3910–3916.
cité pages 61 et 64

[96] Schmidt Am Busch M., Lopes A., Mignon D. et Simonson T. (**2008**) *Computational protein design: software implementation, parameter optimization, and performance of a simple model. Journal of Computational Chemistry*, volume 29, pages 1092–1102.
cité pages 61, 66, 75, 99 et 118

[97] Schmidt am Busch M., Lopes A., Amara N., Bathelt C. et Simonson T. (**2008**) *Testing the Coulomb/Accessible Surface Area solvent model for protein stability, ligand binding, and protein design. BMC Bioinformatics*, volume 9, pages 148–163.
cité pages 61, 65, 66, 92 et 117

[98] Simonson T (**2013**) *Protein:ligand recognition: simple models for electrostatic effects. Curr Pharma Design*, volume 19, pages 4241–4256.
cité pages 61, 66, 92 et 99

[99] Dahiyat B., Gordon D. et Mayo S. (**1997**) *Automated design of the surface positions of protein helices. Protein Science*, volume 6, pages 1333–1337.
cité page 61

[100] Gough, J., Karplus, K., Hughey, R., Chothia et C. (**2001**) *Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. J Mol Biol.*, volume 313 (4), pages 903–19.
cité pages 62, 68 et 100

[101] Wilson D, Madera M, Vogel C, Chothia C et Gough J (**2007**) *The SUPERFAMILY database in 2007: families and functions. Nucl Acids Res*, volume 35, pages D308—D313.
cité pages 62, 68 et 100

[102] David Simoncini, D. Allouche, Simon de Givry, C'eline Delmas, S. Barbe et T. Schiex (**2015**) *Guaranteed Discrete Energy Optimization on Large Protein Design Problems. JCTC*, volume 11, pages 5980–5989.
cité pages 66, 75, 76 et 81

[103] Lee B. et Richards F. (**1971**) *The interpretation of protein structures : estimation of static accessibility. J. Mol. Biol.*, volume 55, pages 379–400.
cité pages 66 et 93

[104] Murphy L. R., Wallqvist A., Levy et R. M. (**2000**) *Simplified amino acid alphabets for protein fold recognition and implications for folding. Prot. Eng.*, volume 13, pages 149–152.
cité page 68

[105] W. L. DeLano (**2002**) *The PyMOL molecular graphics system.* DeLano Scientific, San Carlos, CA, USA.
cité page 70

[106] Voigt C.A., Gordon D.B. et Mayo S.L (**2000**) *Trading accuracy for speed : A quantitative comparison of search algorithms in protein sequence design. J. Mol. Biol.*, volume 299, pages 789–803.
cité pages 79 et 80

[107] Yang W., Wilkins A. L., Ye Y., Liu Z., Li S., Urbauer J., Hellinga H., Kearney A., Der Merwe P. V. et Yang J. (**2005**) *Design of a calcium-binding protein with desired structure in a cell adhesion molecule. J. Am. Chem. Soc.*, volume 127, pages 2085–2093.
cité page 79

[108] Allen B. et Mayo S. (**2010**) *An ecient algorithm for multistate protein design based on FASTER. Journal of Computational Chemistry*, volume 31, pages 904–916.
cité page 79

[109] R. Abagyan et M. Totrov (**1994**) *Biased probability Monte Carlo conformational searches and electrostatic calculations for peptides and proteins. JMB*, volume 235, pages 983–1002.
cité page 79

[110] N. Zhang et C. Zeng (**2008**) *Reference energy extremal optimization: a stochastic search algorithm applied to computational protein design. JCC*, volume 29, pages 1762–1771.
cité page 79

[111] X. Hu, H. Hu, D. N. Beratan et W. Yang (**2010**) *A gradient-directed Monte Carlo approach for protein design. JCC*, volume 31, pages 2164–2168.
cité page 79

[112] H. K. Fung, C. A. Floudas, M. S. Taylor, L. Zhang et D. Morikis (**2008**) *Towards full-sequence de novo protein design with flexible templates for human beta-defensin-2. BPJ*, volume 94, pages 584–599.
cité page 80

[113] Mandell D., Coutsias E. et Kortemme T. (**2009**) *Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling. Natures Methods*, volume 6, pages 551–552.
cité page 80

[114] P. S. Huang, Y. E. Ban, F. Richter, I. Andre, R. Vernon, W. R. Schief et D. Baker (**2011**) *RosettaRemodel: a generalized framework for flexible backbone protein design. PLoS One*, volume 6, pages e24109.
cité page 80

[115] Mark A. Hallen, Pablo Gainza et B. R. Donald (**2015**) *Compact Representation of Continuous Energy Surfaces for More Efficient Protein Design. JCTC*, volume 11, pages 2292–2306.
cité page 80

[116] A. H. Ng et C. D. Snow (**2011**) *Polarizable Protein Packing. JCC*, volume 32, pages 1334–1344.
cité page 80

[117] Stephen D. LuCore, Jacob M. Litman, Kyle T. Powers, Shibo Gao, Ava M. Lynn, William T. A. Tollefson, Timothy D. Fenn, M. Todd Washington et Michael J. Schnieders (**2015**) *Dead-End Elimination with a Polarizable Force Field Repacks PCNA Structures. BPJ*, volume 109, pages 816–826.
cité page 80

[118] T. Simonson, A. Aleksandrov et P. Satpati (**2015**) *Electrostatic free energies in translational GTPases: classic allostery and the rest. BBA Gen. Subj.*, volume 1850, pages 1006–1016.
cité page 81

[119] Harris BZ et Lim WA (**2001**) *Mechanism and role of PDZ domains in signaling complex assembly. J Cell Sci*, volume 114, pages 3219–3231.
cité page 89

[120] Hung AY et Sheng M (**2002**) *PDZ Domains: Structural Modules for Protein Complex Assembly. J Biol Chem*, volume 277, pages 5699–5702.
cité page 89

[121] Tonikian R, Zhang YN, Sazinsky SL, Currell B, Yeh JH, Reva B et et al (**2008**) *A Specificity Map for the PDZ Domain Family. PLoS Biology*, volume 6, pages 2043–2059.
cité page 89

[122] Gfeller D, Butty F, Wierzbicka M, Verschueren E, Vanhee P, Huang H et et al (**2011**) *The multiplespecificity landscape of modular peptide recognition domains. Molec Syst Biol*, volume 7, pages 484.
cité page 89

[123] Subbaiah VK, Kranjec C, Thomas M et Ban L (**2011**) *Structural and thermodynamic analysis of PDZ-ligand interactions. Biochem J*, volume 439, pages 195–205.
cité page 89

[124] Roberts KE, Cushing PR, Boisguerin P, Madden DR et Donald BR (**2012**) *Computational Design of a PDZ Domain Peptide Inhibitor that Rescues CFTR Activity. PLoS Comp Bio*, volume 8, pages e1002477.
cité page 89

[125] Zheng F, Jewell H, Fitzpatrick J, Zhang J, Mierke DF et Grigoryan G (**2015**) *Computational Design of Selective Peptides to Discriminate between Similar PDZ Domains in an Oncogenic Pathway. J Mol Biol*, volume 427, pages 491–510.
cité page 89

[126] Kong Y et Karplus M (**2009**) *Signaling pathways of PDZ2 domain: A molecular dynamics interaction correlation analysis. Proteins*, volume 74, pages 145–154.
cité page 89

[127] McLaughlin Jr RN, Poelwijk FJ, Raman A, Gosal WS et Ranganathan R (**2012**) *The spatial architecture of protein function and adaptation. Nature*, volume 458, pages 859–864.
cité page 89

[128] Melero C, Ollikainen N, Harwood I, Karpiak J et Kortemme T (**2014**) *Quantification of the transferability of a designed protein specificity switch reveals extensive epistasis in molecular recognition. Proc Natl Acad Sci USA*, volume 111, pages 15426–15431.
cité page 89

[129] Reina J, Lacroix E, Hobson SD, Fernandez-Ballester G, Rybin V, Schwab MS et et al (**2002**) *Computer-aided design of a PDZ domain to recognize new target sequences. Nat Struct Mol Biol*, volume 9, pages 621–627.
cité page 89

[130] Smith CA et Kortemme T (**2010**) *Structure-Based Prediction of the Peptide Sequence Space Recognized by Natural and Synthetic PDZ Domains. J Mol Biol*, volume 402, pages 460–474.
cité page 89

[131] Baker D (**2006**) *Prediction and design of macromolecular structures and interactions. Phil Trans R Soc Lond*, volume 361, pages 459–463.
cité pages 89 et 100

[132] Kleinman CL, Rodrigue N, Bonnard C, Philippe H et Lartillot N (**2006**) *A maximum likelihood framework for protein design. BMC Bioinf*, volume 7, pages Art. 326.
cité pages 91 et 92

[133] Druart K, Palmai Z, Omarjee E et Simonson T (**2016**) *Protein:ligand binding free energies: a stringent test for computational protein design. J Comput Chem*, volume 37, pages 404–415.
cité page 93

[134] Pieper U. Eswar N. Marti-Renom M. A. Webb B. Madhusudhan M. S. Eramian D. Shen M. et Sali M (**2006**) *Comparative Protein Structure Modeling With MODELLER. Curr. Prot. Bioinf.*, volume Suppl. 15, pages 5.6.1–5.6.30.
cité page 99

[135] Jaramillo A, Wernisch L, Hery S et Wodak S (**2002**) *Folding free energy function selects nativelike protein sequences in the core but not on the surface. Proc Natl Acad Sci USA*, volume 99, pages 13554–13559.
cité page 103

[136] Druart K., Le Guennec M., Palmai Z. et Simonson T. (**2017**) *Probing the stereospecificity of tyrosyl- and glutaminyl-trna synthetase with molecular dynamics simulations. Journal of Molecular Graphics and Modelling*, volume 71, pages 192–199.
cité page 117

[137] Schmidt am Busch M., Sedano A. et T. Simonson (**2010**) *Computational protein design: validation and possible relevance as a tool for homology searching and fold recognition. PLoS One*, volume 5(5), pages e10410.
cité page 118

[138] Panel N. (**2017**) *Etude computationnelle du domaine PDZ de Tiam1.* Thèse de doctorat de l'Université Paris-Saclay.
cité page 118

Remerciements

Je remercie les membres du jury, Sophie Barbe, Julien Bigot, Alain Denis, Jean-François Gibrat et Yves-Henri Sanejouand pour avoir accepté de lire cette thèse et d'évaluer mon travail. Je remercie particulièrement Thomas Simonson, mon directeur de thèse, pour son encadrement de qualité, son énergie à avancer et pour tout ce que cela m'a déjà apporté. Je remercie Yves Mechulam, le directeur de mon laboratoire, de m'avoir permis d'effectuer ce travail et de m'avoir offert de très bonnes conditions pour le réaliser tout au long de ces dernières années.

Ma thèse s'inscrit dans un projet bien plus large dans lequel beaucoup de personnes ont déjà collaboré. J'ai bénéficié de leurs travaux. Et je remercie Thomas Gaillard, pour la qualité de sa contribution au projet sur laquelle se fondent beaucoup de mes résultats, Nicolas Panel pour notre collaboration sur la famille PDZ et son expertise de Cask et Tiam1, Francesco Villa pour ses tests proteus et son apport du GB, Karen Druart pour nos discussions sur le Monte Carlo, Alexey Aleskandrov pour sa vision générale de la bio-informatique structurale, mais aussi, Anne Lopes, Najette Amara et Audrey Sedano.

J'ai eu la chance de collaborer avec plusieurs autres équipes. Je remercie Julien Bigot de la maison de la simulation pour ses conseils sur l'implémentation du parallélisme. Je remercie Isabelle Dupays et Laurent Leger de l'IDRIS pour leur étude des performances de proteus et Seydou Traoré pour ses conseils d'utilisation de Toulbar2.

université PARIS-SACLAY
ÉCOLE DOCTORALE
INTERFACES
Approches Interdisciplinaires : Fondements, Applications et Innovation

Titre : Computational protein design: un outil pour l'ingénierie des protéines et la biologie synthétique

Mots clés : modélisation moléculaire, conception de protéine par ordinateur, Proteus, Monte Carlo, domaine PDZ

Résumé : Le « Computational protein design » ou CPD est la recherche des séquences d'acides aminés compatibles avec une structure protéique ciblée. L'objectif est de concevoir une fonction nouvelle et/ou d'ajouter un nouveau comportement. Le CPD est en développement dans de notre laboratoire depuis plusieurs années, avec le logiciel Proteus qui a plusieurs succès à son actif. Notre approche utilise un modèle énergétique basé sur la physique et s'appuie sur la différence d'énergie entre l'état plié et l'état déplié de la protéine. Au cours de cette thèse, nous avons enrichi Proteus sur plusieurs points, avec notamment l'ajout d'une méthode d'exploration Monte Carlo avec échange de répliques ou REMC. Nous avons comparé trois méthodes stochastiques pour l'exploration de l'espace de la séquence : le REMC, le Monte Carlo simple et une heuristique conçue pour le CPD, le « Multistart Steepest Descent » ou MSD. Ces comparaisons portent sur neuf protéines de trois familles de structures : SH2, SH3 et PDZ. En utilisant les techniques d'exploration ci-dessus, nous avons été en mesure d'identifier la conformation du minimum global d'énergie ou GMEC pour presque tous les tests dans lesquels jusqu'à 10 positions de la chaîne polypeptidique étaient libres de muter (les autres conservant leurs types natifs). Pour les tests avec 20 positions libres de muter, le GMEC a été identifié dans 2/3 des cas. Globalement, le REMC et le MSD donnent de très bonnes séquences en termes d'énergie, souvent identiques ou très proches du GMEC. Le MSD a obtenu les meilleurs résultats sur les tests à 30 positions mutables. Le REMC avec huit répliques et des paramètres optimisés a donné le plus souvent le meilleur résultat lorsque toutes les positions peuvent muter. De plus, comparé à une énumération exacte des séquences de faible énergie, le REMC fournit un échantillon de séquences de grande diversité.

Dans la seconde partie de ce travail, nous avons testé notre modèle pour la conception de domaines PDZ. Pour l'état plié, nous avons utilisé deux variantes d'un modèle de solvant GB. La première utilise une frontière diélectrique protéine/solvant effective moyenne ; la seconde, plus rigoureuse, utilise une frontière exacte qui fluctue le long de la trajectoire MC. Pour caractériser l'état déplié, nous utilisons un ensemble de potentiels chimiques d'acide aminé ou énergies de références. Ces énergies de références sont déterminées par maximisation d'une fonction de vraisemblance afin de reproduire les fréquences d'acides aminés des domaines PDZ naturels. Les séquences conçues par Proteus ont été comparées aux séquences naturelles. Nos séquences sont globalement similaires aux séquences Pfam, au sens des scores BLOSUM40, avec des scores particulièrement élevés pour les résidus au cœur de la protéine. La variante de GB la plus rigoureuse donne toujours des séquences similaires à des homologues naturels modérément éloignés et l'outil de reconnaissance de plis Superfamily appliqué à ces séquences donne une reconnaissance parfaite. Nos séquences ont également été comparées à celles du logiciel Rosetta. La qualité, selon les mêmes critères que précédemment, est très comparable, mais les séquences Rosetta présentent moins de mutations que les séquences Proteus.

Title: Computational protein design: a tool for protein engineering and synthetic biology

Keywords: molecular modeling, computational protein design, Proteus, Monte Carlo, PDZ domain

Abstract: Computational Protein Design, or CPD is the search for the amino acid sequences compatible with a targeted protein structure. The goal is to design a new function and/or add a new behavior. CPD has been developed in our laboratory for several years, with the software Proteus which has several successes to its credit. Our approach uses a physics-based energy model, and relies on the energy difference between the folded and unfolded states of the protein. During this thesis, we enriched Proteus on several points, including the addition of a Monte Carlo exploration method with Replica Exchange or REMC. We compared extensively three stochastic methods for the exploration of sequence space: REMC, plain Monte Carlo and a heuristic designed for CPD: Multistart Steepest Descent or MSD. These comparisons concerned nine proteins from three structural families: SH2, SH3 and PDZ. Using the exploration techniques above, we were able to identify the Global Minimum Energy Conformation, or GMEC for nearly all the test cases where up to 10 positions of the polypeptide chain were free to mutate (the others retaining their native types). For the tests where 20 positions were free to mutate, the GMEC was identified in 2/3 of the cases. Overall, REMC and MSD give very good sequences in terms of energy, often identical or very close to the GMEC. MSD performed best in the tests with 30 mutating positions. REMC with eight replicas and optimized parameters often gave the best result when all positions could mutate. Moreover, compared to an exact enumeration of the low energy sequences, REMC provided a sample of sequences with a high sequence diversity.

In the second part of this work, we tested our CPD model for PDZ domain design. For the folded state, we used two variants of a GB solvent model. The first used a mean, effective protein/solvent dielectric boundary; the second one, more rigorous, used an exact boundary that flucutated over the MC trajectory. To characterize the unfolded state, we used a set of amino acid chemical potentials or reference energies. These reference energies were determined by maximizing a likelihood function so as to reproduce the amino acid frequencies in natural PDZ domains. The sequences designed by Proteus were compared to the natural sequences. Our sequences are globally similar to the Pfam sequences, in the sense of the BLOSUM40 scores, with especially high scores for the residues in the core of the protein. The more rigorous GB variant always gives sequences similar to moderately distant natural homologues and perfect recognition by the the Superfamily fold recognition tool. Our sequences were also compared to those produced by the Rosetta software. The quality, according to the same criteria as before, was very similar, but the Rosetta sequences exhibit fewer mutations than the Proteus sequences.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France