OTTO-VON-GUERICKE UNIVERSITY MAGDEBURG

Visualisation Toolkit for Contact Density Potentials within Amino Acid Neighbourhoods in Protein Structures

by

Corinna Vehlow

supervised by

Prof. Dr. Bernhard Preim and Dr. Michael Lappe

A diploma thesis submitted in partial fulfillment for the degree of Graduate Engineer of Computational Visualistics

in the

Faculty of Computer Science

Department of Simulation and Graphics

October 2010

Declaration of Authorship

I, Corinna Vehlow, declare that this thesis titled, "Visualisation Toolkit for Contact Density Potentials within Amino Acid Neighbourhoods in Protein Structures" and the work presented in it are my own. I confirm that only the sources and means cited have been used. Parts that are direct quotes or paraphrases are identified as such. The same is true of tables and figures that have been used. This thesis, or any part of it has not previously been submitted for a degree or any other qualification at this University or any other institution.

Signed: _____

Date: _____

OTTO-VON-GUERICKE UNIVERSITY MAGDEBURG

Abstract

Faculty of Computer Science
Department of Simulation and Graphics

Graduate Engineer of Computational Visualistics

by Corinna Vehlow

The prediction of the folding of protein structures is of great value for the investigation of causes for and development of cancer as well as for the purpose of drug design. This thesis will introduce a developed software, which extracts and visualises orientation propensities of amino acids within proteins and discusses how to slip the derived information into the process of structure prediction. Therefore, first of all an introduction into the biological background, particularly the structure of proteins and experimental techniques used to derive a protein's folding, will be given. This will be followed by some related work regarding possible visualisation techniques. After a comprehensive description of the concept and the implemented visualisation, the latter will be evaluated with respect to its usability. Different possibilities, describing how to make use of the derived information, will be discussed in the context of sampling methods and energy functions as well as scoring functions for the process of structure prediction. Finally, a conclusion will be drawn.

Acknowledgements

Sincere thanks to Prof. Dr. B. Preim, who was supervising me and helping with words and deeds during the my studies, especially during my final project up to the present day and supported me with lots of advice.

Furthermore, I would like to thank Dr. M. Lappe, who was guiding me through the project of this thesis on site from the very first beginning. Thanks for all the advice and help in all respects and for stimulating my research. Many thanks as well to H. Stehr and M. Winkelmann, who often gave me some inspiring input and support. I would also like to thank the rest of the "Bioinformatics Research Group" of M. Lappe for their help during the validation process, particularly for the time they spent on evaluation procedure. Many thanks to Lars Petzold, who provided me with some illustrations.

Finally many thanks to my family, especially my parents, and my friends, who are always backing me with what I am doing and have confidence into my plans. Without those people I would not have become the kind of person I am.

Contents

Abbreviations

AA	**A**mino **A**cid
BB	**B**ack**b**one
CASP	**C**ritical **A**ssessment of Protein **S**tructure **P**rediction
CATH	**C**lassification of
CCD	**C**harged **C**oupled **D**evice
COSY	**Co**rrelation **S**pectroscop**y**
DBSCAN	**D**ensity **B**ased **S**patial **C**lustering of **A**pplications with **N**oise
DNA	**D**eoxyribo**n**ucleic **a**cid
DOI	**D**egree **o**f **I**nterest
EDM	**E**lectron **D**ensity **M**ap
GDT	**G**lobal **D**istance **T**est
LOS	**L**og-**O**dds-**S**core
MAD	**M**ultiwavelength **A**nomalous **D**iffraction
MIR	**M**ultiple **I**somorphous **R**eplacement
MMDS	**M**etric **M**ulti-**d**imensional **S**caling
NBH	**N**eigh**b**our**h**ood
NMR	**N**uclear **M**agnetic **R**esonance
NOE	**N**uclear **O**verhauser **E**ffect
PDB	**P**rotein **D**ata **B**ank
RCC	**R**ank **C**orrelation **C**oefficient
RF	**R**adio **F**requency
RMSD	**R**oot **M**ean **S**quare **D**eviation
SC	**S**ide**c**hain
SCOP	**S**tructural **C**lassification **o**f **P**roteins
SS	**S**econdary **S**tructure
SSE	**S**econdary **S**tructure **E**lement
SHS	**S**pherical **H**armonic **S**ynthesis

Symbols

λ	angle (latitude)	$^\circ$
ϕ	angle (longitude)	$^\circ$
ψ	angle	$^\circ$
θ	angle	$^\circ$
Å	Ångström	
C	carbon atom	
H	hydrogen atom	
O	oxygen atom	
N	nitrogen atom	

Chapter 1

Introduction

Protein structure prediction is of high importance for various scientific areas. Chemists use information about the binding sites for the purpose of drug design. In contrast, biochemists are much more interested in the three-dimensional structure of proteins, which gives insights into its particular *cellular function*. The position of certain sequence features of a protein within the 3D structure is of substantial and practical value. The knowledge about the spatial location of certain residues within the structure can be used for site directed mutagenesis. The rate at which protein structures are solved has increased, whereas to predict the 3D structure from sequence still remains a challenge. Up to date more than 60.000 high-resolution protein structures are available in public protein data banks (PDB). For less than 1% of the known protein sequences the 3D structure was determined experimentally, i.e. with the help of X-ray crystallography or nuclear magnetic resonance (NMR) spectroscopy.

There exist different approaches for structure prediction. As lots of proteins are quite similar regarding their sequence, one approach is to use template PDB structures and alignment methods, like homology modelling or protein threading. Besides comparative protein modelling, ab initio or de novo protein modelling are used, especially for those sequences, where no template PDB structures could be found. The aim of protein modelling methods is to identify the most plausible structure, which is usually done via energy functions and potentials of mean force.

Many different approaches to derive potentials for interaction determination and structure prediction can be found in the literature. Residues of a protein's primary sequence interact with each other, if they are connected via hydrogen bonds and situated close to each other, i.e. their distance deceeds a given threshold. Most of those methods, used to determine interactions, rely on the assumption that known X-ray or NMR resolved protein structures represent classical equilibrium states. They are used to derive some kinds of standards and principles. Statistical methods, like the Boltzmann device, are used quite often (e.g. by [Sip90]) to obtain *distance dependent mean force potentials*. Many studies, like that of Tobi et al. [TE00, TSLE00] use only distance constraints.

Such distance constraints for the residues of the protein define, how far or close certain amino acids of the chain are situated from each other. These constraints are further used for structure prediction, as they give information on how the residues might be situated to each other. Nevertheless, such constraints do not allow to determine structures unambiguously. This includes e.g. the occurrence of spiegelmers, i.e. mirror-inverted 3D structures of an amino acid chain.
Another structural, quite important parameter that could be used to determine how chains fold to proteins is the relative packing of side chains within proteins. The way how a chain folds and how neighbouring residues of a chain are placed to one another can be described through two angles (phi and psi). Those dihedral angles pose an important determinant of local geometry, i.e. secondary structure, as well as three-dimensional topology, i.e. tertiary structure. Bahar and Jernigan were the first, who captured the existence of relative orientation probabilities [BJ96]. To show that the *preferred orientation of side chains depends* on the nature of the residue, they used a simple body fixed coordinate system. Their results implied that statistical potentials are sensitive to orientation.
The backbone (BB) atoms of amino acids (AA) itself are relatively fixed within a plane, whereas the attached side chain (SC) can have different orientations. Neighboured BBs of the amino acid sequence are connected over a certain angle to each other. The way, how certain AAs are placed to one another, is defined by the dihedral angle pair mentioned above. Those angles for all involved contacts define the tertiary structure of the protein. The aim in terms of protein fold recognition and structure prediction is to obtain accurate residue-dependent potentials of spatial arrangement. Besides those angles between BB elements, also the rotamers of attached side chains are of interest. Rotamers are favoured conformations, i.e. orientations of hydrogen bonds between atoms of side chains that occur more frequently as they are energetically more favourable [IntroProteinStructure].
Summarising, the analysis of contacts and side chain packing, offering distance and *orientation constraints* supports the development of knowledge-based potentials. To overcome ambiguities in structure prediction, which occur when using distance constraints alone, the definition of at least some orientation constraints between residues might help.

1.1 Objective

The aim of this project is to develop a graphical support to detect spatially preferred orientations of specific residue contacts. Based on the data analysis of available resolved protein structures, the potential spatial distribution of residue contacts can be extracted. This includes the alignment of contacts along an orientation and translation invariant framework.
The visualisations should highlight preferred dihedral angles and thus help to define

several orientation constraints for a subset of the residue contacts. Together with the distance constraints those orientation constraints could be handed over to the reconstruction method, which will hopefully produce more precise models of the 3D structure. In the best case, the problem of spiegelmers will be removed, thus giving the correct of the two possible structures. The definition of orientation constraints might also help to analyse the side chain packing, i.e. how good the side chains are intermeshed. In the long run, it could expand the conditions for scoring functions that are used to assess models, produced by different prediction methods.

The following paragraphs will give an overview about the biological background of this thesis. The structure of proteins will be discussed, after giving an introduction about amino acids and its properties. This will be followed by a short résumé about the two experimental techniques, X-ray and NMR spectroscopy, used for structure prediction. Afterwards, related work in the field of visualisation of protein structures will be summarised and discussed. This will be complemented through the discussion of various general visualisation techniques, which could be used for the aim of this thesis. Based on the introduced visualisation possibilities the concept will be presented and discussed. Afterwards the implementation will be presented and evaluated, followed by a discussion about future perspectives and potential improvements.

Chapter 7

Conclusion

The determination of the three-dimensional structure based on the sequence information is of great value for a number of reasons. The knowledge about the structure, which defines a protein's function, gives insights into the molecular basis of different diseases, e.g. cancer or Alzheimer's. Further, it introduces advanced possibilities for the purpose of drug design.
The speed at which protein structures are predicted lacks behind the determination of new sequences. This is because currently used experimental techniques, like NMR and X-ray crystallography, are quite difficult and time consuming and can not be applied to every protein of interest. The prediction of 3D structures solely based on the primary sequence still poses a problem considering accuracy. There are different approaches based on homology modelling, which can produce models with an RMSD of about $1 - 2$Å for models with high sequence identity of about 70%. At least, such techniques can be used to produce hypotheses about a protein's function. Besides the template selection and alignment of the target with a protein template, the process of homology modelling also includes the model construction and further model assessment. The latter can be performed via the methods statistical potentials or physics-based energy calculation. Such statistical potentials comprise empirical methods based on observed residue-residue frequencies among known protein structures. On the basis of such statistics various constraints can be derived, which can be used to decrease the amount of possible conformations or at least to rate possible states. Nevertheless, there is still a high degree of flexibility concerning the 3D structure, the distance constraints do not determine the structure bijectively. Rather, there exists an ensemble of views that is energetically possible.
That is where the extraction and further visualisation of statistically preferred orientations of specific residues comes into play. These geometric propensities can be used to support the process of model assessment. The visualisation of these propensities clarifies, that there exist clear tendencies for residues to be located at some predominant position with respect to each other. It can therefore directly illustrate how good or bad

a model is. For this purpose, the gained information only needs to be evaluated and integrated in an appropriate fashion.
Above all, the displays can be utilised to estimate the quality of a predicted model via analysing how well it fits into the statistical background information. This of course requires some user interaction and embodies a rather subjective evaluation. On the other hand, the gained knowledge about preferred residue dependent orientations can be used to calculate quantitative scores about the quality of a predicted model. This presupposes an adequate and well-defined scoring function.
Besides the assessment of the quality of a predicted model, there still lies a potential in the use of geometric orientation propensities for the definition of orientation constraints. These can be formulated as dihedral angles which could be taken into account by sampling methods used to reduce the amount of possible conformations during the phase of model assessment, thereby extending these methods towards a more funded basis. The definition of dihedrals requires a high degree of interaction through the user. This is justified by the necessity to evaluate and analyse several contacts in the context of their residue environments (NBHs). At the same time, the implemented options for clustering, cluster and histogram analysis can help within the decision making process. Especially the extraction of predominantly occurring residue types within the clusters supports the determination of the preferred orientations of residues with respect to each other.
To what extent there exist a potential in improving the predicted model, via extending the set of constraints for sampling methods, might be better assessable after the development of a scoring (energy) function and the comparison of this with a reference score like the GDT score. For this purpose, the scoring function should evaluate solely in how far the position of $jRes$ with respect to $iRes$ suits the statistical distribution of nodes, i.e. whether or not it coincides with a cluster that preferably contains residues of the same type. In case that a developed scoring function, based on such statistical information, correlates well with other established scores, we can assume that the derived geometric orientation constraints will yield an increase in accuracy.
The determination and extraction of statistical residue and orientation dependent probabilities facilitates to gain further insights into the mechanisms of protein folding. These information can be used and integrated into the process of homology-based structure modelling in general and the phase of model assessment in particular.

Bibliography

[BBM03] Vladimir Batagelj, Vladimir Batagelj, and Andrej Mrvar. Pajek: analysis and visualization of large networks. In *Graph Drawing Software*, pages 77–103. Springer, 2003.

[BEPW08] Klaus Backhaus, Bernd Erichson, Wulff Plinke, and Rolf Weiber. *Multivariate Analysemethoden: Eine anwendungsorientierte Einführung.* Springer, Berlin, 12., vollständig überarbeitete auflage. edition, 2008.

[BEW95] Richard A. Becker, Stephen G. Eick, and Allan R. Wilks. Visualizing network data. *IEEE Transactions on Visualization and Computer Graphics*, 1(1):16–28, 1995.

[BJ96] I. Bahar and R.L. Jernigan. Coordination geometry of nonbonded residues in globular proteins. *Folding & design*, 1(5):357–70, Jan 1996.

[Boe10] R. Boehm. Die ganze kartennetzentwurfslehre kurzgefasst, Mai 2010.

[BS95] L.M. Bugayevskiy and J.P. Snyder. *Map Projections: A reference manual.* CRC Press, first edition edition, June 1995.

[BSP+93] Eric A. Bier, Maureen C. Stone, Ken Pier, William Buxton, and Tony D. DeRose. Toolglass and magic lenses: the see-through interface. In *SIGGRAPH '93: Proceedings of the 20th annual conference on Computer graphics and interactive techniques*, pages 73–80, New York, NY, USA, 1993. ACM.

[BST03a] B.-J.J. Breitkreutz, C. Stark, and M. Tyers. Osprey: a network visualization system. *Genome biology*, 4(3), 2003.

[BST03b] N.-V. Buchete, J.E. Straub, and D. Thirumalai. Anisotropic coarse-grained statistical potentials improve the ability to identify nativelike protein structures. *J Chem Phys*, 118(16):7658–7671, Jan 2003.

[BST04a] N.-V. Buchete, J.E. Straub, and D. Thirumalai. Continuous anisotropic representation of coarse-grained potentials for proteins by spherical harmonics synthesis. *J Mol Graph Model*, 22(5):441–50, May 2004.

[BST04b] N.-V. Buchete, J.E. Straub, and D. Thirumalai. Development of novel statistical potentials for protein fold recognition. *Curr Opin Struct Biol*, 14(2):225–32, Apr 2004.

[BT99] C.-I. Branden and J. Tooze. *Introduction to Protein Structure: Second Edition*. Garland Publishing, second edition, January 1999.

[BW04] Ulrik Brandes and Dorothea Wagner. Netzwerkvisualisierung. *it - Information Technology*, 46(3):129–134, 2004.

[BWK+01] A. Vilanova Bartrol, R. Wegenkittl, A. Knig, E. Grller, E. Sorantin, and Tiani Medgraph. Virtual colon flattening, 2001.

[CMS99] Stuart K. Card, J. D. Mackinlay, and Ben Shneiderman. *Readings in Information Visualization: Using Vision to Think*. Academic Press, London, 1999.

[DAK09] Swagatam Das, Ajith Abraham, and Amit Konar. *Metaheuristic Clustering*. Springer Publishing Company, Incorporated, 2009.

[Dan00] Peter H. Dana. Map projections, 10 2000.

[Dav05] da vinci. http://www.informatik.uni-bremen.de/uDrawGraph/en/index.html, 2005.

[DeL] W.L. DeLano. *The PyMOL User's Manual*. DeLano Scientific, Palo Alto, CA, USA.

[DS03] Richard P Dum and Peter L Strick. An unfolded map of the cerebellar dentate nucleus and its projections to the cerebral cortex. *Journal for Neurophysiology*, 89(1):634–639, 2003.

[Fur86] G. Furnas. Generalized fisheye views. *CHI '86: Proceedings of the SIGCHI conference on Human factors in computing systems*, Apr 1986.

[Fur09] C.A. Furuti. Cartographical map projections, July 2009.

[Gra00] Graphlet toolkit. http://www.infosun.fim.uni-passau.de/Graphlet/, 2000.

[HAK00] Steven Haker, Sigurd Angenent, and Ron Kikinis. Nondistorting flattening maps and the 3d visualization of colon ct images. *IEEE Trans. on Medical Imaging*, 19:665–670, 2000.

[HCvW07] D. Holten, B. Cornelissen, and J.J. van Wijk. Trace visualization using hierarchical edge bundles and massive sequence views. pages 47 –54, jun. 2007.

[HGQ+06] Wei Hong, Xianfeng Gu, Feng Qiu, Miao Jin, and Arie Kaufman. Conformal virtual colon flattening. In *SPM '06: Proceedings of the 2006 ACM symposium on Solid and physical modeling*, pages 85–93, New York, NY, USA, 2006. ACM.

[Hol06] Danny Holten. Hierarchical edge bundles: Visualization of adjacency relations in hierarchical data. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):741–748, 2006.

[JAV07] A.N. Jha, G.K. Ananthasuresh, and S. Vishveshwara. Protein sequence design based on the topology of the native state structure. *J Theor Biol*, 248(1):81–90, Sep 2007.

[JH04] C. Johnson and C. Hansen. *Visualization Handbook*. Academic Press, Inc., Orlando, FL, USA, 2004.

[JSJ+84] Richard E. Rosenfield Jr, Stanley M. Swanson, Edgar F. Meyer Jr, Horace L. Carrell, and Peter Murray-Rust. Mapping the atomic environment of functional groups: turning 3d scatter plots into pseudo-density contours. *Journal of Molecular Graphics*, 2(2):43 – 46, 1984.

[KLD+10] Kristian W. Kaufmann, Gordon H. Lemmon, Samuel L. DeLuca, Jonathan H. Sheehan, and Jens Meiler. Practically useful: What the rosetta protein modeling suite can do for you. *Biochemistry*, 49(14):2987–2998, 2010. PMID: 20235548.

[LBF+09] M. Lappe, G. Bagler, I. Filippis, H. Stehr, J.M. Duarte, and R. Sathyapriya. Designing evolvable libraries using multi-body potentials. *Current Opinion in Biotechnology*, 20(4):437 – 446, 2009. Protein technologies / Systems and synthetic biology.

[Maz09] Riccardo Mazza. *Introduction to Information Visualization*. Springer, 1 edition, March 2009.

[MBHC95] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol*, 247(4):536–540, Apr 1995.

[MFK+09] J. Moult, K. Fidelis, A. Kryshtafovych, B. Rost, and A. Tramontano. Critical assessment of methods of protein structure predictionround viii. volume 77, pages 1–4. Wiley Subscription Services, Inc., A Wiley Company, August 2009.

[NGB+09] M. Neugebauer, R. Gasteiger, O. Beuing, V.Diehl, M. Skalej, and B. Preim. Map Displays for the Analysis of Scalar Data on Cerebral Aneurysm Surfaces. In *Computer Graphics Forum (EuroVis)*, volume 28 (3), pages 895–902, Berlin, 10.-12. Juni 2009.

[OGF+10] S.I. O'Donoghue, D.S. Goodsell, A.S. Frangakis, F. Jossinet, R.A.Laskowski, M. Nilges, H.R. Saibil, A. Schafferhans, R.C. Wade, E.Westhof, and A.J. Olson. Visualization of macromolecular structures. *Nature methods*, 7(3 Suppl), March 2010.

[OGH+06] Steffen Oeltze, Frank Grothues, Anja Hennemuth, Anja Ku, and Bernhard Preim. Integrated Visualization of Morphologic and Perfusion Data for the Analysis of Coronary Artery Disease. In *IEEE/Eurographics Symposium on Visualization*, Informatik aktuell, pages 131–138. Springer, 2006.

[OMJ+97] Ca Orengo, Ad Michie, S Jones, Dt Jones, and Mb Swindells. Cath:a hierarchic classification of protein domain structures. *Structure*, 5:1093–1108, 1997.

[Pea90] Frederic Pearson. *Map Projections: Theory and Applications*. CRC Press, second edition edition, March 1990.

[Phi69] D.C. Phillips. The development of crystallographic enzymology. *Biochemical Society Symposium*, 30:11–28, Nov 1969.

[PWS08] G. Pavlopoulos, A.L. Wegener, and R. Schneider. A survey of visualization tools for biological network analysis. *BioData Mining*, 1(1):12+, November 2008.

[RA08] S. Rakshit and G.K. Ananthasuresh. An amino acid map of inter-residue contact energies using metric multi-dimensional scaling. *J Theor Biol*, 250(2):291–7, Jan 2008.

[Ran00] M. Randić. Condensed representation of dna primary sequences. *Journal of Chemical Information and Computer Sciences*, 40(1):50–56, Jan 2000.

[Rho06] Gale Rhodes. *Crystallography made crystal clear - A guide for users of macromolecular models*. Academic Press, Complementary Science Series, third edition, February 2006.

[RRS63] G.N. Ramachandran, C. Ramakrishnan, and V. Sasisekharan. Stereochemistry of polypeptide chain configurations. *Molecular Biology*, 7(1):95–99, Jan 1963.

[RWS+10] Christian Rieder, Andreas Weihusen, Christian Schumann, Stephan Zidowitz, and Heinz-Otto Peitgen. Visual support for interactive post-interventional assessment of radiofrequency ablation therapy. volume 29, pages 1093–1102. Blackwell Publishing Ltd, June 2010.

[SDS+09] R. Sathyapriya, J.M. Duarte, H. Stehr, I. Filippis, and M. Lappe. Defining an essence of structure determining residue contacts in proteins. *PLoS Comput Biol*, 5(12):e1000584, 12 2009.

[Sip90] Manfred J. Sippl. Calculation of conformational ensembles from potentials of mean force. an approach to the knowledge-based prediction of local structures in globular proteins. *J Mol Biol*, 213(4):859–83, Jun 1990.

[SMO+03] P Shannon, A Markiel, O Ozier, N S Baliga, J T Wang, D Ramage, N Amin, B Schwikowski, and T Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*, 13(11):2498–2504, November 2003.

[Sny93] J.P. Snyder. *Flattening the Earth: Two Thousand Years of Map Projections*. University of Chicago Press, 1993.

[SS06] Min-yi Shen and Andrej Sali. Statistical potential for assessment and prediction of protein structures. *Protein Science*, 15(11):2507–2524, November 2006.

[TE00] D. Tobi and R. Elber. Distance-dependent, pair potential for protein folding: results from linear optimization. *Proteins*, 41(1):40–6, Oct 2000.

[Tid05] J. Tidwell. *Designing Interfaces : Patterns for Effective Interaction Design*. O'Reilly Media, Inc., November 2005.

[TSLE00] D. Tobi, G. Shafran, N. Linial, and R. Elber. On the design and analysis of protein folding potentials. *Proteins*, 40(1):71–85, Jul 2000.

[VS93] G. Vriend and C. Sander. Quality-control of protein models - directional atomic contact analysis. *J Appl Crystallogr*, 26:47–60, Jan 1993.

[WD03] Guoli Wang and L. Dunbrack. Pisces: a protein sequence culling server. *Bioinformatics*, 19(12):1589–1591, Aug 2003.

[Wei10] E.W. Weisstein. Wolframmathworld the web's most extensive mathematics resource, April 2010.

[Wes74] Westermann, editor. *Diercke Weltatlas*. Georg Westermann Verlag, Braunschweig, 180. auflage (92. auflage der neubearbeitung) edition, 1974.

[WLYY03] Y.H. Wu, A.W.C. Liew, H. Yan, and M.S. Yang. Db-curve: a novel 2d method of dna sequence visualization and representation. *Chem Phys Lett*, 367(1-2):170–176, Jan 2003.

[XB05] Lei Xie and Philip E. Bourne. Functional coverage of the human genome by existing structures, structural genomics targets, and homology models. *PLoS Comput. Biol*, 1(3):31, August 2005.

[YST02] Qihe. Yang, John P. Snyder, and Waldo R. Tobler. *Map projection transformation : principles and applications*. Taylor & Francis, 2002.

[ZB07] M. Zvelebil and J.O. Baum. *Understanding Bioinformatics*. Garland Science, first edition, August 2007.