References

Ahmad, S. & Sarai, A. (2005). PSSM-based prediction of DNA binding sites in proteins. *BMC Bioinformatics*, **6**, 33. 26

Alber, T., Gilbert, W.A., Ponzi, D.R. & Petsko, G.A. (1983). The role of mobility in the substrate binding and catalytic machinery of enzymes. *Ciba Found Symp*, **93**, 4–24. 111

Amico, M., Finelli, M., Rossi, I., Zauli, A., Elofsson, A., Viklund, H., von Heijne, G., Jones, D., Krogh, A., Fariselli, P., Martelli, P.L. & Casadio, R. (2006). PONGO: a web server for multiple predictions of all-alpha transmembrane proteins. *Nucleic Acids Res*, **34**, W169–W172. 137

Andersen, J.S., Lyon, C.E., Fox, A.H., Leung, A.K.L., Lam, Y.W., Steen, H., Mann, M. & Lamond, A.I. (2002). Directed proteomic analysis of the human nucleolus. *Curr Biol*, **12**, 1–11. 113

Andersen, J.S., Lam, Y.W., Leung, A.K.L., Ong, S.E., Lyon, C.E., Lamond, A.I. & Mann, M. (2005). Nucleolar proteome dynamics. *Nature*, **433**, 77–83. 113

Bailey, T.L. & Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol*, **2**, 28–36. 3

Bailey, T.L. & Elkan, C. (1995). The value of prior knowledge in discovering motifs with meme. *Proc Int Conf Intell Syst Mol Biol*, **3**, 21–29. 18, 27, 64

Bairoch, A. & Apweiler, R. (1996). The SWISS-PROT protein sequence data bank and its new supplement TREMBL. *Nucleic Acids Res*, **24**, 21–25. 12, 40

Bairoch, A. & Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res*, **28**, 45–48. 12

Bannai, H., Tamada, Y., Maruyama, O., Nakai, K. & Miyano, S. (2002). Extensive feature detection of N-terminal protein sorting signals. *Bioinformatics*, **18**, 298–305. 7, 9

Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E.L.L., Studholme, D.J., Yeats, C. & Eddy, S.R. (2004). The Pfam protein families database. *Nucleic Acids Res*, **32**, D138–D141. 11, 25, 58

Bendtsen, J.D., Jensen, L.J., Blom, N., Heijne, G.V. & Brunak, S. (2004a). Feature-based prediction of non-classical and leaderless protein secretion. *Protein Eng Des Sel*, **17**, 349–356. 65

Bendtsen, J.D., Nielsen, H., von Heijne, G. & Brunak, S. (2004b). Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol*, **340**, 783–795. 7, 8

Bhasin, M. & Raghava, G.P.S. (2004). ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST. *Nucleic Acids Res*, **32**, W414–W419. 7, 10

Bhasin, M., Garg, A. & Raghava, G.P.S. (2005). PSLpred: prediction of subcellular localization of bacterial proteins. *Bioinformatics*, **21**, 2522–2524. 10

BioJava (2007). http://www.biojava.org. 27, 75, 148, 150

Brameier, M., Krings, A. & MacCallum, R.M. (2007). NucPred–predicting nuclear localization of proteins. *Bioinformatics*, **23**, 1159–1160. 110

Burgard, A.P., Moore, G.L. & Maranas, C.D. (2001). Review of the TEIRESIAS-based tools of the IBM Bioinformatics and Pattern Discovery Group. *Metab Eng*, **3**, 285–288. 26

Burges, C.J.C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, **2**, 121–167. 23, 24

Burley, S.K. & Petsko, G.A. (1985). Aromatic-aromatic interaction: a mechanism of protein structure stabilization. *Science*, **229**, 23–28. 132

Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M.C., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., Wells, C., Kodz-

ius, R., Shimokawa, K., Bajic, V.B., Brenner, S.E., Batalov, S., Forrest, A.R.R., Zavolan, M., Davis, M.J., Wilming, L.G., Aidinis, V., Allen, J.E., Ambesi-Impiombato, A., Apweiler, R., Aturaliya, R.N., Bailey, T.L., Bansal, M., Baxter, L., Beisel, K.W., Bersano, T., Bono, H., Chalk, A.M., Chiu, K.P., Choudhary, V., Christoffels, A., Clutterbuck, D.R., Crowe, M.L., Dalla, E., Dalrymple, B.P., de Bono, B., Gatta, G.D., di Bernardo, D., Down, T., Engstrom, P., Fagiolini, M., Faulkner, G., Fletcher, C.F., Fukushima, T., Furuno, M., Futaki, S., Gariboldi, M., Georgii-Hemming, P., Gingeras, T.R., Gojobori, T., Green, R.E., Gustincich, S., Harbers, M., Hayashi, Y., Hensch, T.K., Hirokawa, N., Hill, D., Huminiecki, L., Iacono, M., Ikeo, K., Iwama, A., Ishikawa, T., Jakt, M., Kanapin, A., Katoh, M., Kawasawa, Y., Kelso, J., Kitamura, H., Kitano, H., Kollias, G., Krishnan, S.P.T., Kruger, A., Kummerfeld, S.K., Kurochkin, I.V., Lareau, L.F., Lazarevic, D., Lipovich, L., Liu, J., Liuni, S., McWilliam, S., Babu, M.M., Madera, M., Marchionni, L., Matsuda, H., Matsuzawa, S., Miki, H., Mignone, F., Miyake, S., Morris, K., Mottagui-Tabar, S., Mulder, N., Nakano, N., Nakauchi, H., Ng, P., Nilsson, R., Nishiguchi, S., Nishikawa, S., Nori, F., Ohara, O., Okazaki, Y., Orlando, V., Pang, K.C., Pavan, W.J., Pavesi, G., Pesole, G., Petrovsky, N., Piazza, S., Reed, J., Reid, J.F., Ring, B.Z., Ringwald, M., Rost, B., Ruan, Y., Salzberg, S.L., Sandelin, A., Schneider, C., Schnbach, C., Sekiguchi, K., Semple, C.A.M., Seno, S., Sessa, L., Sheng, Y.,

Shibata, Y., Shimada, H., Shimada, K., Silva, D., Sinclair, B., Sperling, S., Stupka, E., Sugiura, K., Sultana, R., Takenaka, Y., Taki, K., Tammoja, K., Tan, S.L., Tang, S., Taylor, M.S., Tegner, J., Teichmann, S.A., Ueda, H.R., van Nimwegen, E., Verardo, R., Wei, C.L., Yagi, K., Yamanishi, H., Zabarovsky, E., Zhu, S., Zimmer, A., Hide, W., Bult, C., Grimmond, S.M., Teasdale, R.D., Liu, E.T., Brusic, V., Quackenbush, J., Wahlestedt, C., Mattick, J.S., Hume, D.A., Kai, C., Sasaki, D., Tomaru, Y., Fukuda, S., Kanamori-Katayama, M., Suzuki, M., Aoki, J., Arakawa, T., Iida, J., Imamura, K., Itoh, M., Kato, T., Kawaji, H., Kawagashira, N., Kawashima, T., Kojima, M., Kondo, S., Konno, H., Nakano, K., Ninomiya, N., Nishio, T., Okada, M., Plessy, C., Shibata, K., Shiraki, T., Suzuki, S., Tagami, M., Waki, K., Watahiki, A., Okamura-Oho, Y., Suzuki, H., Kawai, J., Hayashizaki, Y., Consortium, F.A.N.T.O.M., Group, R.I.K.E.N.G.E.R. & Group), G.S.G.G.N.P.C. (2005). The transcriptional landscape of the mammalian genome. *Science*, **309**, 1559–1563. 114

Chang, C.C. & Lin, C.J. (2001). LIBSVM: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/∼cjlin/libsvm. 23, 24, 79, 108, 118

Chothia, C. & Lesk, A.M. (1986). The relation between the divergence of sequence and structure in proteins. *EMBO J*, **5**, 823–826. 13

Claros, M.G. & von Heijne, G. (1994). TopPred II: an improved software for membrane protein structure predictions. *Comput Appl Biosci*, **10**, 685–686.

68, 137

Coeytaux, K. & Poupon, A. (2005). Prediction of unfolded segments in a protein sequence based on amino acid composition. *Bioinformatics*, **21**, 1891–1900. 112

Cokol, M., Nair, R. & Rost, B. (2000). Finding nuclear localization signals. *EMBO Rep*, **1**, 411–415. 7, 8, 110

Comon, P. (1994). Independent component analysis, a new concept? *Signal Process.*, **36**, 287–314. 19

Cozzetto, D., Matteo, A.D. & Tramontano, A. (2005). Ten years of predictions ... and counting. *FEBS J*, **272**, 881–882. 112, 159

Cozzetto, D., Giorgetti, A., Raimondo, D. & Tramontano, A. (2007). The evaluation of protein structure prediction results. *Mol Biotechnol*. 112, 159

Dang, C.V. & Lee, W.M. (1989). Nuclear and nucleolar targeting sequences of c-erb-a, c-myb, N-myc, p53, HSP70, and HIV tat proteins. *J Biol Chem*, **264**, 18019–18023. 108, 120

Dempster, A., Laird, N. & Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J Royal Statistical Society, Series B*, **39**. 18

Dingwall, C. & Laskey, R.A. (1991). Nuclear targeting sequences– a consensus? *Trends Biochem Sci*, **16**, 478–481. 73

Doğruel, M., Down, T. & Hubbard, T. (2008). NestedMICA as an ab initio protein motif discovery tool. *BMC Bioinformatics*, **9**, 19. 3, 25, 64, 69, 115

Down, T.A. & Hubbard, T.J.P. (2002). Computational detection and location of transcription start sites in mammalian genomic DNA. *Genome Res*, **12**, 458–461. 3, 27, 64, 73, 74, 76

Down, T.A. & Hubbard, T.J.P. (2004). What can we learn from noncoding regions of similarity between genomes? *BMC Bioinformatics*, **5**, 131. 74

Down, T.A. & Hubbard, T.J.P. (2005). NestedMICA: sensitive inference of over-represented motifs in nucleic acid sequence. *Nucleic Acids Res*, **33**, 1445–1453. 2, 27, 30, 32, 33, 34, 36, 64

Dunker, A.K., Obradovic, Z., Romero, P., Garner, E.C. & Brown, C.J. (2000). Intrinsic protein disorder in complete genomes. *Genome Inform Ser Workshop Genome Inform*, **11**, 161–171. 111, 168

Durbin, R., Eddy, S.R., Krogh, A. & Mitchison, G. (1999). *Biological Sequence Analysis : Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press. 16, 18, 30

Eck, R.V. & Dayhoff, M.O. (1966). *Atlas of Protein Sequence and Structure*, vol. 3. National Biomedical Research Foundation, Silver Spring, Maryland. 68

Edwards, R.J., Davey, N.E. & Shields, D.C. (2007). SLiMFinder: a probabilistic method for identifying over-represented, convergently evolved, short linear motifs in proteins. *PLoS ONE*, **2**, e967. 26

Emanuelsson, O., Nielsen, H. & von Heijne, G. (1999). ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites. *Protein Sci*, **8**, 978–984. 7, 8, 66, 85, 136

Emanuelsson, O., Nielsen, H., Brunak, S. & von Heijne, G. (2000). Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J Mol Biol*, **300**, 1005–1016. 7, 40, 56, 59

Emanuelsson, O., Brunak, S., von Heijne, G. & Nielsen, H. (2007). Locating proteins in the cell using TargetP, SignalP and related tools. *Nat Protoc*, **2**, 953–971. 8

Endres, M., Neupert, W. & Brunner, M. (1999). Transport of the ADP/ATP carrier of mitochondria from the TOM complex to the TIM22.54 complex. *EMBO J*, **18**, 3214–3221. 65

Favorov, A.V., Gelfand, M.S., Gerasimova, A.V., Ravcheev, D.A., Mironov, A.A. & Makeev, V.J. (2005). A Gibbs sampler for identification of symmetrically structured, spaced DNA motifs with improved estimation of the signal length. *Bioinformatics*, **21**, 2240–2245. 27

Fink, J.L., Aturaliya, R.N., Davis, M.J., Zhang, F., Hanson, K., Teasdale, M.S., Kai, C., Kawai, J., Carninci, P., Hayashizaki, Y. & Teasdale, R.D. (2006). LOCATE: a mouse protein subcellular localization database. *Nucleic Acids Res*, **34**, D213–D217. 12, 114

Gao, Y. & Mehta, K. (2007). N-linked glycosylation of CD38 is required for its structure stabilization but not for membrane localization. *Mol Cell Biochem*, **295**, 1–7. 67

Gardy, J.L., Spencer, C., Wang, K., Ester, M., Tusnády, G.E., Simon, I., Hua, S., deFays, K., Lambert, C., Nakai, K. & Brinkman, F.S.L. (2003). PSORT-B: Improving protein subcellular localization prediction for Gram-negative bacteria. *Nucleic Acids Res*, **31**, 3613–3617. 9

Gardy, J.L., Laird, M.R., Chen, F., Rey, S., Walsh, C.J., Ester, M. & Brinkman, F.S.L. (2005). PSORTb v.2.0: expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis. *Bioinformatics*, **21**, 617–623. 9, 13

Garner, Romero, Dunker, Brown & Obradovic (1999). Predicting binding regions within disordered proteins. *Genome Informatics*, **10**, 41–50. 112

Goldfarb, D.S., Garipy, J., Schoolnik, G. & Kornberg, R.D. (1986). Synthetic peptides as nuclear localization signals. *Nature*, **322**, 641–644. 108, 116

Gould, S.G., Keller, G.A. & Subramani, S. (1987). Identification of a peroxisomal targeting signal at the carboxy terminus of firefly luciferase. *J Cell Biol*, **105**, 2923–2931. 65

Gould, S.J., Keller, G.A., Hosken, N., Wilkinson, J. & Subramani, S. (1989). A conserved tripeptide sorts proteins to peroxisomes. *J Cell Biol*, **108**, 1657–1664. 65

Gribskov, M., Mclachlan, A.D. & Eisenberg, D. (1987). Profile analysis: detection of distantly related proteins. 17

Guan, J.L., Machamer, C.E. & Rose, J.K. (1985). Glycosylation allows cell-surface transport of an anchored secretory protein. *Cell*, **42**, 489–496. 67

Guda, C. & Subramaniam, S. (2005). pTARGET [CORRECTED] a new method for predicting protein subcellular localization in eukaryotes. *Bioinformatics*, **21**, 3963–3969. 7, 10, 13, 36, 70, 71

Hannink, M. & Donoghue, D.J. (1986). Cell surface expression of membrane-anchored v-sis gene products: glycosylation is not required for cell surface transport. *J Cell Biol*, **103**, 2311–2322. 67

Henikoff, S. & Henikoff, J.G. (1992). Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A*, **89**, 10915–10919. 68

Hertz, G.Z. & Stormo, G.D. (1999). Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, **15**, 563–577. 26

Hinsby, A.M., Kiemer, L., Karlberg, E.O., Lage, K., Fausbll, A., Juncker, A.S., Andersen, J.S., Mann, M. & Brunak, S. (2006). A wiring of the human nucleolus. *Mol Cell*, **22**, 285–295. 114

Hirokawa, T., Boon-Chieng, S. & Mitaku, S. (1998). SOSUI: classification and secondary structure prediction system for membrane proteins. *Bioinformatics*, **14**, 378–379. 68, 137

Hobohm, U., Scharf, M., Schneider, R. & Sander, C. (1992). Selection of representative protein data sets. *Protein Sci*, **1**, 409–417. 13, 15, 41, 56

Höglund, A., Dönnes, P., Blum, T., Adolph, H.W. & Kohlbacher, O. (2006). MultiLoc: prediction of protein subcellular localization using N-terminal targeting sequences, sequence motifs and amino acid composition. *Bioinformatics*, **22**, 1158–1165. 7, 10, 11, 13, 41, 70, 71, 97, 98, 99, 180

Horton, P. & Nakai, K. (1997). Better prediction of protein cellular localization sites with the k nearest neighbors classifier. *Proc Int Conf Intell Syst Mol Biol*, **5**, 147–152. 7, 9

Horton, P., Park, K.J., Obayashi, T., Fujita, N., Harada, H., Adams-Collier, C.J. & Nakai, K. (2007). WoLF PSORT: protein localization predictor. *Nucleic Acids Res*, **35**, W585–W587. 7, 9

Hsu, C.W. & Lin, C.J. (2002). A simple decomposition method for support vector machines. *Machine Learning*, **46**, 291–314. 24, 79

Hua, S. & Sun, Z. (2001). Support vector machine approach for protein subcellular localization prediction. *Bioinformatics*, **17**, 721–728. 7, 10

Huber, R. (1979). Conformational flexibility in protein molecules. *Nature*, **280**, 538–539. 111

Hughes, J.D., Estep, P.W., Tavazoie, S. & Church, G.M. (2000). Computational identification of cis-regulatory elements associated with groups of

functionally related genes in Saccharomyces cerevisiae. *J Mol Biol*, **296**, 1205–1214. 27

Hulo, N., Bairoch, A., Bulliard, V., Cerutti, L., Castro, E.D., Langendijk-Genevaux, P.S., Pagni, M. & Sigrist, C.J.A. (2006). The PROSITE database. *Nucleic Acids Res*, **34**, D227–D230. 11, 25, 40

Hwang, S., Gou, Z. & Kuznetsov, I.B. (2007). DP-Bind: a web server for sequence-based prediction of DNA-binding residues in DNA-binding proteins. *Bioinformatics*, **23**, 634–636. 26

Jin, Y. & Dunbrack, R.L. (2005). Assessment of disorder predictions in CASP6. *Proteins*, **61 Suppl 7**, 167–175. 113

Joachims, T. (1999). Making large-scale support vector machine learning practical. 169–184, MIT Press. 23, 79

Jones, D.T. (2007). Improving the accuracy of transmembrane protein topology prediction using evolutionary information. *Bioinformatics*, **23**, 538–544. 137, 159

Jones, D.T., Taylor, W.R. & Thornton, J.M. (1994). A model recognition approach to the prediction of all-helical membrane protein structure and topology. *Biochemistry*, **33**, 3038–3049. 142

Juncker, A.S., Willenbrock, H., Heijne, G.V., Brunak, S., Nielsen, H. & Krogh, A. (2003). Prediction of lipoprotein signal peptides in Gram-negative bacteria. *Protein Sci*, **12**, 1652–1662. 7, 8

Käll, L., Krogh, A. & Sonnhammer, E.L.L. (2004). A combined transmembrane topology and signal peptide prediction method. *J Mol Biol*, **338**, 1027–1036. 137, 139, 159, 160

Käll, L., Krogh, A. & Sonnhammer, E.L.L. (2007). Advantages of combined transmembrane topology and signal peptide prediction–the Phobius web server. *Nucleic Acids Res*, **35**, W429–W432. 138, 164

Kaplan, H.A., Welply, J.K. & Lennarz, W.J. (1987). Oligosaccharyl transferase: the central enzyme in the pathway of glycoprotein assembly. *Biochim Biophys Acta*, **906**, 161–173. 67

Kelley, L.P. & Kinsella, B.T. (2003). The role of N-linked glycosylation in determining the surface expression, G protein interaction and effector coupling of the alpha isoform of the human thromboxane A(2) receptor. *Biochim Biophys Acta*, **1621**, 192–203. 67

Kiemer, L., Bendtsen, J.D. & Blom, N. (2005). NetAcet: prediction of N-terminal acetylation sites. *Bioinformatics*, **21**, 1269–1270. 38

Kissinger, C.R., Parge, H.E., Knighton, D.R., Lewis, C.T., Pelletier, L.A., Tempczyk, A., Kalish, V.J., Tucker, K.D., Showalter, R.E. & Moomaw, E.W. (1995). Crystal structures of human calcineurin and the human FKBP12-FK506-calcineurin complex. *Nature*, **378**, 641–644. 132

Klein, P., Kanehisa, M. & DeLisi, C. (1984). Prediction of protein function from sequence properties. Discriminant analysis of a data base. *Biochim Biophys Acta*, **787**, 221–226. 68

Krogh, A. (2002). *www.binf.ku.dk/∼krogh/docs/labeled_fasta_format.html*. 148

Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E.L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol*, **305**, 567–580. 6, 68, 79, 123, 124, 134, 137, 138, 140

Kuznetsov, I.B., Gou, Z., Li, R. & Hwang, S. (2006). Using evolutionary and structural information to predict DNA-binding sites on DNA-binding proteins. *Proteins*, **64**, 19–27. 27

la Cour, T., Kiemer, L., Mølgaard, A., Gupta, R., Skriver, K. & Brunak, S. (2004). Analysis and prediction of leucine-rich nuclear export signals. *Protein Eng Des Sel*, **17**, 527–536. 110

Lao, D.M., Okuno, T. & Shimizu, T. (2002). Evaluating transmembrane topology prediction methods for the effect of signal peptide in topology prediction. *In Silico Biol*, **2**, 485–494. 69, 137

Leung, A.K.L., Andersen, J.S., Mann, M. & Lamond, A.I. (2003). Bioinformatic analysis of the nucleolus. *Biochem J*, **376**, 553–569. 111, 125

Li, Romero, Rani, Dunker & Obradovic (1999). Predicting protein disorder for N-, C-, and internal regions. *Genome Informatics*, **10**, 30–40. 112

Li, H. & Bingham, P.M. (1991). Arginine/serine-rich domains of the su(wa) and tra RNA processing regulators target proteins to a subnuclear compartment implicated in splicing. *Cell*, **67**, 335–342. 108

Li, W. & Godzik, A. (2006). CD-HIT: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659. 14, 15, 36, 70, 71, 115

Li, W., Jaroszewski, L. & Godzik, A. (2001). Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics*, **17**, 282–283. 14

Li, W., Jaroszewski, L. & Godzik, A. (2002). Tolerating some redundancy significantly speeds up clustering of large protein databases. *Bioinformatics*, **18**, 77–82. 14, 15

Linding, R., Jensen, L.J., Diella, F., Bork, P., Gibson, T.J. & Russell, R.B. (2003a). Protein disorder prediction: implications for structural proteomics. *Structure*, **11**, 1453–1459. 112

Linding, R., Russell, R.B., Neduva, V. & Gibson, T.J. (2003b). GlobPlot: Exploring protein sequences for globularity and disorder. *Nucleic Acids Res*, **31**, 3701–3708. 112

Lobley, A., Swindells, M.B., Orengo, C.A. & Jones, D.T. (2007). Inferring function using patterns of native disorder in proteins. *PLoS Comput Biol*, **3**, e162. 102, 112, 168

Lu, Z., Szafron, D., Greiner, R., Lu, P., Wishart, D.S., Poulin, B., Anvik, J., Macdonell, C. & Eisner, R. (2004). Predicting subcellular localization of proteins using machine-learned classifiers. *Bioinformatics*, **20**, 547–556. 7, 10

Lumb, M.J., Purdue, P.E. & Danpure, C.J. (1994). Molecular evolution of alanine/glyoxylate aminotransferase 1 intracellular targeting. Analysis of the feline gene. *Eur J Biochem*, **221**, 53–62. 104

Machamer, C.E., Florkiewicz, R.Z. & Rose, J.K. (1985). A single N-linked oligosaccharide at either of the two normal sites is sufficient for transport of vesicular stomatitis virus G protein to the cell surface. *Mol Cell Biol*, **5**, 3074–3083. 67

Maeda, N., Kasukawa, T., Oyama, R., Gough, J., Frith, M., Engstrm, P.G., Lenhard, B., Aturaliya, R.N., Batalov, S., Beisel, K.W., Bult, C.J., Fletcher, C.F., Forrest, A.R.R., Furuno, M., Hill, D., Itoh, M., Kanamori-Katayama, M., Katayama, S., Katoh, M., Kawashima, T., Quackenbush, J., Ravasi, T., Ring, B.Z., Shibata, K., Sugiura, K., Takenaka, Y., Teasdale, R.D., Wells, C.A., Zhu, Y., Kai, C., Kawai, J., Hume, D.A., Carninci, P. & Hayashizaki, Y. (2006). Transcript annotation in FANTOM3: mouse gene catalog based on physical cDNAs. *PLoS Genet*, **2**, e62. 114

Maeda, Y., Hisatake, K., Kondo, T., Hanada, K., Song, C.Z., Nishimura, T. & Muramatsu, M. (1992). Mouse rRNA gene transcription

factor mUBF requires both HMG-box1 and an acidic tail for nucleolar accumulation: molecular analysis of the nucleolar targeting mechanism. *EMBO J*, **11**, 3695–3704. 108

Martelli, P.L., Fariselli, P. & Casadio, R. (2003). An ENSEMBLE machine learning approach for the prediction of all-alpha membrane proteins. *Bioinformatics*, **19 Suppl 1**, i205–i211. 137

Matsuda, K., Zheng, J., Du, G.G., Klcker, N., Madison, L.D. & Dallos, P. (2004). N-linked glycosylation sites of the motor protein prestin: effects on membrane targeting and electrophysiological function. *J Neurochem*, **89**, 928–938. 67

Matthews, B.W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta*, **405**, 442–451. 38, 47, 80

Milstein, C., Brownlee, G.G., Harrison, T.M. & Mathews, M.B. (1972). A possible precursor of immunoglobulin light chains. *Nat New Biol*, **239**, 117–120. 66

Mohrmann, K., van Eijndhoven, M.A.J., Schinkel, A.H. & Schellens, J.H.M. (2005). Absence of N-linked glycosylation does not affect plasma membrane localization of breast cancer resistance protein (BCRP/ABCG2). *Cancer Chemother Pharmacol*, **56**, 344–350. 67

Motley, A., Lumb, M.J., Oatey, P.B., Jennings, P.R., Zoysa, P.A.D., Wanders, R.J., Tabak, H.F. & Danpure, C.J. (1995). Mammalian ala-

nine/glyoxylate aminotransferase 1 is imported into peroxisomes via the PTS1 translocation pathway. Increased degeneracy and context specificity of the mammalian PTS1 motif and implications for the peroxisome-to-mitochondrion mistargeting of AGT in primary hyperoxaluria type 1. *J Cell Biol*, **131**, 95–109. 104

Müller, S., Croning, M.D. & Apweiler, R. (2001). Evaluation of methods for the prediction of membrane spanning regions. *Bioinformatics*, **17**, 646–653. 69, 137

Munro, S. (1995). An investigation of the role of transmembrane domains in Golgi protein retention. *EMBO J*, **14**, 4695–4704. 95

Nair, R. & Rost, B. (2002a). Inferring sub-cellular localization through automated lexical analysis. *Bioinformatics*, **18 Suppl 1**, S78–S86. 7, 9

Nair, R. & Rost, B. (2002b). Sequence conserved for subcellular localization. *Protein Sci*, **11**, 2836–2847. 7, 9

Nair, R. & Rost, B. (2004). LOCnet and LOCtarget: sub-cellular localization for structural genomics targets. *Nucleic Acids Res*, **32**, W517–W521. 9, 110

Nair, R., Carter, P. & Rost, B. (2003). NLSdb: database of nuclear localization signals. *Nucleic Acids Res*, **31**, 397–399. 8, 11, 110

Nakai, K. & Horton, P. (1999). PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends Biochem Sci*, **24**, 34–36. 7, 9, 110

Nakai, K. & Kanehisa, M. (1991). Expert system for predicting protein localization sites in gram-negative bacteria. *Proteins*, **11**, 95–110. 7, 9, 68

Nakai, K. & Kanehisa, M. (1992). A knowledge base for predicting protein localization sites in eukaryotic cells. *Genomics*, **14**, 897–911. 9

Neduva, V. & Russell, R.B. (2006). DILIMOT: discovery of linear motifs in proteins. *Nucleic Acids Res*, **34**, W350–W355. 26, 58

Ng, P., Nagarajan, N., Jones, N. & Keich, U. (2006). Apples to apples: improving the performance of motif finders and their significance analysis in the twilight zone. *Bioinformatics*, **22**, e393–e401. 34

Nickel, W. (2003). The mystery of nonclassical protein secretion. A current view on cargo proteins and potential export routes. *Eur J Biochem*, **270**, 2109–2119. 66

Nielsen, H. & Krogh, A. (1998). Prediction of signal peptides and signal anchors by a hidden Markov model. *Proc Int Conf Intell Syst Mol Biol*, **6**, 122–130. 160

Nielsen, H., Engelbrecht, J., Brunak, S. & von Heijne, G. (1997a). Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng*, **10**, 1–6. 160

Nielsen, H., Engelbrecht, J., Brunak, S. & von Heijne, G. (1997b). A neural network method for identification of prokaryotic and eukaryotic signal

peptides and prediction of their cleavage sites. *Int J Neural Syst*, **8**, 581–599. 8, 138

Nielsen, H., Brunak, S. & von Heijne, G. (1999). Machine learning approaches for the prediction of signal peptides and other protein sorting signals. *Protein Eng*, **12**, 3–9. 7

Obenauer, J.C., Cantley, L.C. & Yaffe, M.B. (2003). Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs. *Nucleic Acids Res*, **31**, 3635–3641. 25

Oda, T., Miyajima, H., Suzuki, Y. & Ichiyama, A. (1987). Nucleotide sequence of the cDNA encoding the precursor for mitochondrial serine:pyruvate aminotransferase of rat liver. *Eur J Biochem*, **168**, 537–542. 104

Osumi, T., Tsukamoto, T., Hata, S., Yokota, S., Miura, S., Fujiki, Y., Hijikata, M., Miyazawa, S. & Hashimoto, T. (1991). Amino-terminal presequence of the precursor of peroxisomal 3-ketoacyl-coa thiolase is a cleavable signal peptide for peroxisomal targeting. *Biochem Biophys Res Commun*, **181**, 947–954. 65

Park, K.J. & Kanehisa, M. (2003). Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs. *Bioinformatics*, **19**, 1656–1663. 7, 10, 13

Pavesi, G., Mereghetti, P., Mauri, G. & Pesole, G. (2004). Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. *Nucleic Acids Res*, **32**, W199–W203. 27

Pelham, H.R. (1995). Sorting and retrieval between the endoplasmic reticulum and Golgi apparatus. *Curr Opin Cell Biol*, **7**, 530–535. 66, 88

Pierleoni, A., Martelli, P.L., Fariselli, P. & Casadio, R. (2006). BaCelLo: a balanced subcellular localization predictor. *Bioinformatics*, **22**, e408–e416. 7, 10

Prilusky, J., Felder, C.E., Zeev-Ben-Mordehai, T., Rydberg, E.H., Man, O., Beckmann, J.S., Silman, I. & Sussman, J.L. (2005). FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded. *Bioinformatics*, **21**, 3435–3438. 112

Puntervoll, P., Linding, R., Gemnd, C., Chabanis-Davidson, S., Mattingsdal, M., Cameron, S., Martin, D.M.A., Ausiello, G., Brannetti, B., Costantini, A., Ferr, F., Maselli, V., Via, A., Cesareni, G., Diella, F., Superti-Furga, G., Wyrwicz, L., Ramu, C., McGuigan, C., Gudavalli, R., Letunic, I., Bork, P., Rychlewski, L., Kster, B., Helmer-Citterich, M., Hunter, W.N., Aasland, R. & Gibson, T.J. (2003). ELM server: A new resource for investigating short functional sites in modular eukaryotic proteins. *Nucleic Acids Res*, **31**, 3625–3630. 25

Purdue, P.E., Lumb, M.J. & Danpure, C.J. (1992). Molecular evolution of alanine/glyoxylate aminotransferase 1 intracellular targeting. Analysis of the marmoset and rabbit genes. *Eur J Biochem*, **207**, 757–766. 104

Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N., Apweiler, R. & Lopez, R. (2005). InterProScan: protein domains identifier. *Nucleic Acids Res*, **33**, W116–W120. 58

Rabiner, L.R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, **77**, 257–286. 16

Radivojac, P., Obradovic, Z., Brown, C.J. & Dunker, A.K. (2003). Prediction of boundaries between intrinsically ordered and disordered protein regions. *Pac Symp Biocomput*, 216–227. 112

Radivojac, P., Obradovic, Z., Smith, D.K., Zhu, G., Vucetic, S., Brown, C.J., Lawson, J.D. & Dunker, A.K. (2004). Protein flexibility and intrinsic disorder. *Protein Sci*, **13**, 71–80. 112

Ramadass, A.S. (2005). *Computational detection of regulatory signals in human genome sequence*. Ph.D. thesis, University of Cambridge. 74, 75

Reinhardt, A. & Hubbard, T. (1998). Using neural networks for prediction of the subcellular location of proteins. *Nucleic Acids Res*, **26**, 2230–2236. 68

Richardson, J.S. & Richardson, D.C. (1988). Amino acid preferences for specific locations at the ends of alpha helices. *Science*, **240**, 1648–1652. 142

Rigoutsos, I. & Floratos, A. (1998). Combinatorial pattern discovery in biological sequences: The TEIRESIAS algorithm. *Bioinformatics*, **14**, 55–67. 26, 58

Robbins, J., Dilworth, S.M., Laskey, R.A. & Dingwall, C. (1991). Two interdependent basic domains in nucleoplasmin nuclear targeting sequence: identification of a class of bipartite nuclear targeting sequence. *Cell*, **64**, 615–623. 73

Romero, P., Obradovic, Z. & Dunker, A.K. (2004). Natively disordered proteins: functions and predictions. *Appl Bioinformatics*, **3**, 105–113. 112

Sander, C. & Schneider, R. (1991). Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins*, **9**, 56–68. 14

Scherl, A., Cout, Y., Don, C., Call, A., Kindbeiter, K., Sanchez, J.C., Greco, A., Hochstrasser, D. & Diaz, J.J. (2002). Functional proteomic analysis of human nucleolus. *Mol Biol Cell*, **13**, 4100–4109. 113

Schreiber, V., Molinete, M., Boeuf, H., de Murcia, G. & de Murcia, J.M. (1992). The human poly(ADP-ribose) polymerase nuclear localization signal is a bipartite element functionally separate from DNA binding and catalytic activity. *EMBO J*, **11**, 3263–3269. 73

Schultz, J., Milpetz, F., Bork, P. & Ponting, C.P. (1998). SMART, a simple modular architecture research tool: identification of signaling domains. *Proc Natl Acad Sci U S A*, **95**, 5857–5864. 58

Sinha, S. & Tompa, M. (2003). YMF: A program for discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucleic Acids Res*, **31**, 3586–3588. 27

Skilling, J. (2004). Nested Sampling. In R. Fischer, R. Preuss & U.V. Toussaint, eds., *American Institute of Physics Conference Series*, 395–405. 20, 27, 28, 64, 69

Smith, G.B. (1987). Preface to S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images". 562–563, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA. 18

Smola, A. & Scholkopf, B. (1998). A tutorial on support vector regression. Tech. rep. 23

Soro, S. & Tramontano, A. (2005). The prediction of protein function at CASP6. *Proteins*, **61 Suppl 7**, 201–213. 112, 159

Sprenger, J., Fink, J.L. & Teasdale, R.D. (2006). Evaluation and comparison of mammalian subcellular localization prediction methods. *BMC Bioinformatics*, **7 Suppl 5**, S3. 11

Swinkels, B.W., Gould, S.J., Bodnar, A.G., Rachubinski, R.A. & Subramani, S. (1991). A novel, cleavable peroxisomal targeting signal at the amino-terminus of the rat 3-ketoacyl-coa thiolase. *EMBO J*, **10**, 3255–3262. 65

Szafron, D., Lu, P., Greiner, R., Wishart, D.S., Poulin, B., Eisner, R., Lu, Z., Anvik, J., Macdonell, C., Fyshe, A. & Meeuwis, D. (2004). Proteome Analyst: custom predictions with explanations in a web-based tool for high-throughput proteome annotations. *Nucleic Acids Res*, **32**, W365–W371. 13

Takada, Y., Kaneko, N., Esumi, H., Purdue, P.E. & Danpure, C.J. (1990). Human peroxisomal L-alanine: glyoxylate aminotransferase. Evolutionary loss of a mitochondrial targeting signal by point mutation of the initiation codon. *Biochem J*, **268**, 517–520. 104

Thomas, P.D. & Dill, K.A. (1996). An iterative method for extracting energy-like quantities from protein structures. *Proc Natl Acad Sci U S A*, **93**, 11628–11633. 77, 78, 95, 97

Tipping, M.E. (2001). Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, **1**, 211–244. 75

Tompa, M., Li, N., Bailey, T.L., Church, G.M., Moor, B.D., Eskin, E., Favorov, A.V., Frith, M.C., Fu, Y., Kent, W.J., Makeev, V.J., Mironov, A.A., Noble, W.S., Pavesi, G., Pesole, G., Rgnier, M., Simonis, N., Sinha, S., Thijs, G., van Helden, J., Vandenbogaert, M., Weng, Z., Workman, C., Ye, C. & Zhu, Z. (2005). Assessing computational tools for the discovery of transcription factor binding sites. *Nat Biotechnol*, **23**, 137–144. 27

Tusnády, G.E. & Simon, I. (2001). The HMMTOP transmembrane topology prediction server. *Bioinformatics*, **17**, 849–850. 68, 137

Valencia, A. (2005). Protein refinement: a new challenge for CASP in its 10th anniversary. *Bioinformatics*, **21**, 277. 112, 159

Vapnik, V. & Lerner, A. (1963). Pattern recognition using generalized portrait method. *Automation and Remote Control*, **24**, 774–780, 1963. 22

Viklund, H. & Elofsson, A. (2004). Best alpha-helical transmembrane protein topology predictions are achieved using hidden Markov models and evolutionary information. *Protein Sci*, **13**, 1908–1917. 137

von Heijne, G. (1986). A new method for predicting signal sequence cleavage sites. *Nucleic Acids Res*, **14**, 4683–4690. 82, 137

von Heijne, G. (1990). The signal peptide. *J Membr Biol*, **115**, 195–201. 66, 136

Ward, J.J., Sodhi, J.S., McGuffin, L.J., Buxton, B.F. & Jones, D.T. (2004). Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J Mol Biol*, **337**, 635–645. 112

Weinreb, P.H., Zhen, W., Poon, A.W., Conway, K.A. & Lansbury, P.T. (1996). NACP, a protein implicated in Alzheimer's disease and learning, is natively unfolded. *Biochemistry*, **35**, 13709–13715. 111

Wiedemann, N., Pfanner, N. & Ryan, M.T. (2001). The three modules of ADP/ATP carrier cooperate in receptor recruitment and translocation into mitochondria. *EMBO J*, **20**, 951–960. 65

Workman, C.T. & Stormo, G.D. (2000). ANN-Spec: a method for discovering transcription factor binding sites with improved specificity. *Pac Symp Biocomput*, 467–478. 27

Wright, P.E. & Dyson, H.J. (1999). Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J Mol Biol*, **293**, 321–331. 112, 168

Wu, L.F., Chanal, A. & Rodrigue, A. (2000). Membrane targeting and translocation of bacterial hydrogenases. *Arch Microbiol*, **173**, 319–324. 103

Yan, K., Khoshnoodi, J., Ruotsalainen, V. & Tryggvason, K. (2002). N-linked glycosylation is critical for the plasma membrane localization of nephrin. *J Am Soc Nephrol*, **13**, 1385–1389. 67

Yang, Z.R., Thomson, R., McNeil, P. & Esnouf, R.M. (2005). RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. *Bioinformatics*, **21**, 3369–3376. 5, 102, 112, 114, 118

Yu, C.S., Lin, C.J. & Hwang, J.K. (2004). Predicting subcellular localization of proteins for gram-negative bacteria by support vector machines based on n-peptide compositions. *Protein Sci*, **13**, 1402–1406. 7, 10

Yuan, Z. & Teasdale, R.D. (2002). Prediction of Golgi Type II membrane proteins based on their transmembrane domains. *Bioinformatics*, **18**, 1109–1115. 66

Zaslavsky, E. & Singh, M. (2006). A combinatorial optimization approach for diverse motif finding applications. *Algorithms Mol Biol*, **1**, 13. 26