4.6 Summary

We predicted AMPs based on models from different families across various taxa rather than using generalized AMP models. Thomas et. al. created a generalized AMP models to predict AMPs. One limitation of their models are not specific and not accurate. On the other hand, the specific models based on AMPs families is more robust and in addition to predicting AMPs with high accuracy, but also classifies them into specific subclasses such as cecropin, defensin, α-defensin etc.

A large-scale test on all of the currently sequenced and publicly available genomes would be useful to ascertain the robust of the methods used to create the webserver. Establishing the possibility that there are more AMPs through *in*-silico as compared to those discovered in laboratories, can provide an additional sense of direction for the wet lab scientist by testing a few predicted AMPs that have high confidence level for their activity.

The webserver will be useful to scan the ongoing genomes for potential AMPs in insects such as *Anopheles gambiae*, *Glossina morsitans*, *Phlebotomus logipalpis*, *Culex quinquefasciates* and *Anopheles funestus*. These insects are vectors that cause diseases such as trypanoiasis, leishmaniasis, yellow fever and malaria.

One limitation of this study is the lack of enough experimentally validated AMPS that hinders the creation of AMP family models. The other limitation is the small number of sequences used in the target and null databases. The *q*-value and PEP measures depends on the size of the database. The larger the number of sequences in the database that you search, the greater the number of false positives, hence more accurate statistical measure. In future the whole insect proteins from UniProt is intended to be used as the target database in order to generate the null sequences.

Chapter 5

Conclusion and future work

This chapter presents the usage of various computational methods to mine knowledge from the antimicrobial peptides (AMPs) dataset. The main objective of this thesis has been to create AMPs model in order to predict new AMPs. The main contributions of the thesis is broken into three subsections as well as their limitations. In the first subsection, the database of antimicrobial peptides is discussed. In the second subsection, the prediction of AMPs using support vector machines is presented. The section section is on the webserver. Finally, the direction for future work is presented.

UNIVERSITY of the
WESTERN CAPE

5.1 Research contribution and limitations

5.1.1 Antimicrobial peptide database

Databases are useful resource for mining and exploration of antimicrobial peptides, allowing users to query complex biological questions and analysis of data. In this thesis, a comprehensive database of antimicrobial peptides called DAMPD was created. DAMPD is a manually curated database populated with 1232 experimentally validated AMPs entries for both prokaryotic and eukaryotic sources. The procedure for creating the database involves data extraction using keywords and data curation.

The creation of DAMPD database was the first step towards a systematic analysis of AMPs. The DAMPD database was successfully developed and is freely accessible for academic and non-profit users at http://apps.sanbi.ac.za/dampd. The DAMPD database contains both search and analytical tools that

ease in search and analysis of biological query. In particular, classification of AMPs using profile hidden Markov model has been implemented. The profiles created can be used to classify new AMP families into known AMP families. HMM profiles were created for AMPs based on prior knowledge of the AMP families.

5.1.2 Classification of AMPs using support vector machines

Data modeling is usually a crucial step in data mining and yield ground for prediction purposes. The curated data in DAMPD was used to create AMPS models in various taxa. In chapter 3, an SVM-based machine learning approach coupled with optimization methods have been implemented to aid in classification of AMPs into their respective AMPs families. Global optimization methods such as grid search, pattern search and derivative-free simulated annealing were used to select the hyperparameters of SVM classifier. PS-SVM was the best hybrid method based on classification accuracy.

5.1.3 Creation of haemotophagous antimicrobial peptide predictor webserver

A webserver to predict haemotophagous insect AMPs into their respective families was created. The webserver is freely accessible at http://apps.sanbi.ac.za/Happ. This resource is useful to predict AMPs in ongoing genomes.

UNIVERSITY of the
WESTERN CAPE

Some of the future work include

- enriching the database with additional annotation such as information on promoter region and transcription factors for an AMP. The mode of action of AMPs will be added.
- using string kernels such as profile kernel, spectrum kernel and mismatch kernel instead of amino acid composition and physiochemical properties.
- incorporate feature selection in addition to parameter selection of SVM.
- predict AMPs once the genomes for haemotophagous insect are completed.
- use of modified pattern search method that uses perturbed coordinate directions rather than the spanning direction used in PS.

Appendix A

Supplementary material for Chapter 2

ClustalW results page

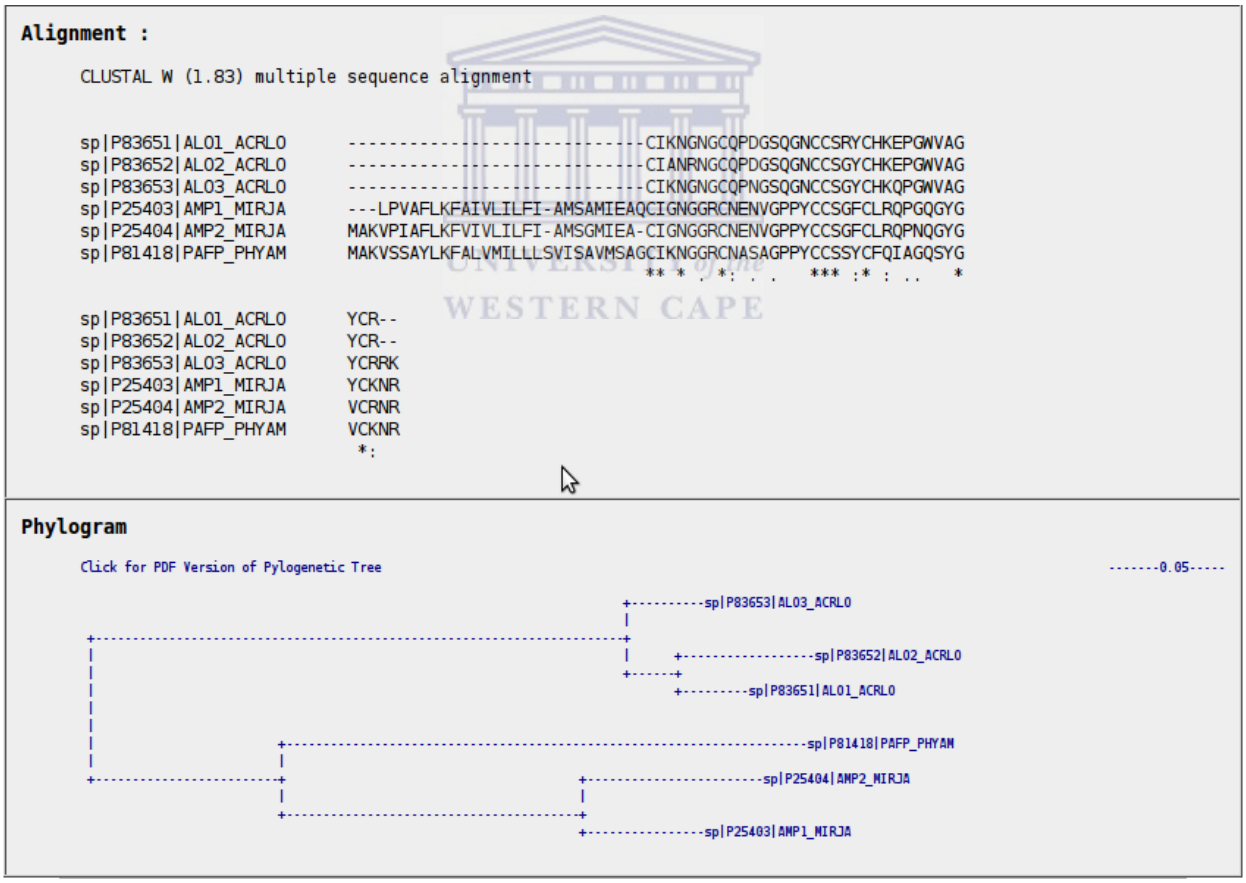


Figure A.1: ClustalW results of AMP family page

HMMER results page

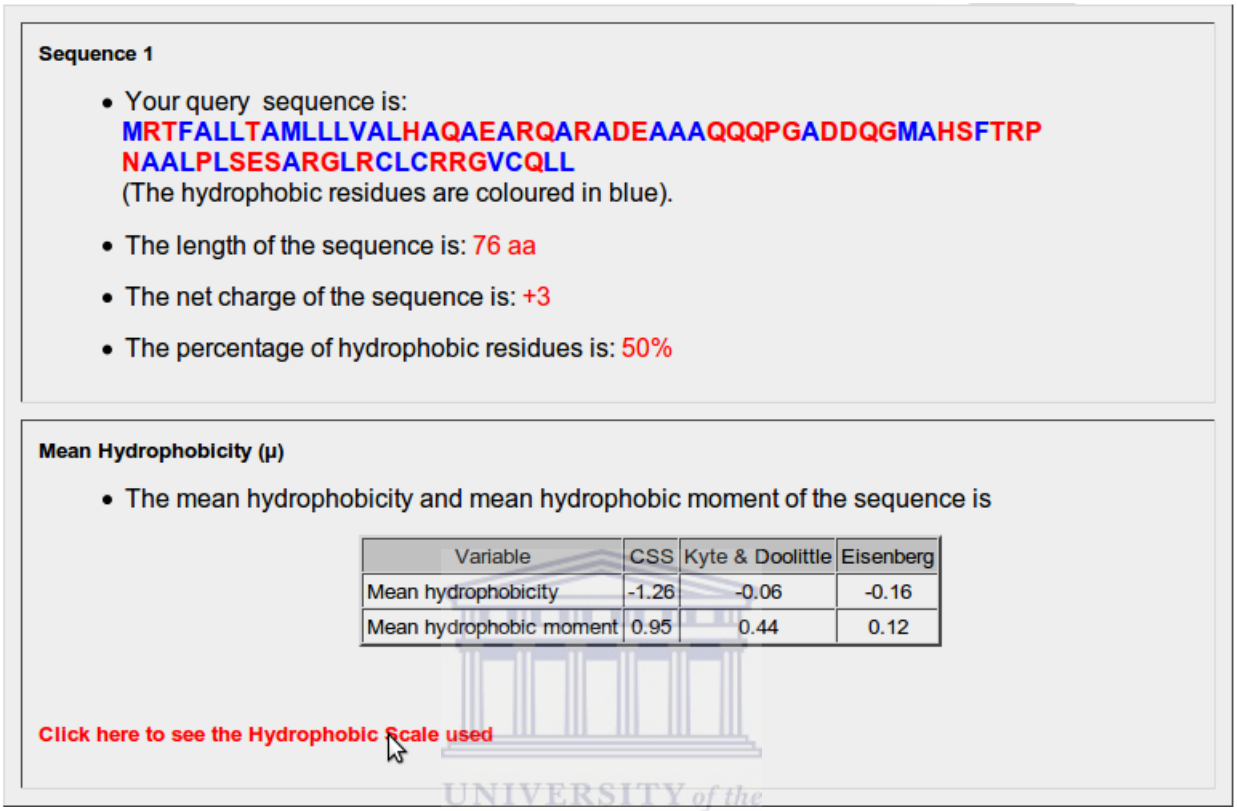
Alignment

```
hmmsearch - search a sequence database with a profile HMM
HMMER 2.3.2 (Oct 2003)
Copyright (C) 1992-2003 HHMI/Washington University School of Medicine
Freely distributed under the GNU General Public License (GPL)
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
HMM file:                 /var/www/dampd/tmp/12225_030411.hmm [12225_030411]
Sequence database:        /var/www/dampd/tmp/12225_030411.query
per-sequence score cutoff:  [none]
per-domain score cutoff:    [none]
per-sequence Eval cutoff:   <= 10
per-domain Eval cutoff:     [none]
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

Query HMM:   12225_030411
Accession:   [none]
Description: [none]
  [HMM has been calibrated; E-values are empirical estimates]

Scores for complete sequences (score includes all domains):
Sequence          Description                        Score    E-value  N
--------          -----------                        -----    ------- ---
sp|Q62715|DEF2_RAT Neutrophil antibiotic peptide NP-2  186.0    1.1e-56   1

Parsed for domains:
Sequence          Domain  seq-f seq-t    hmm-f hmm-t      score  E-value
--------          ------- ----- -----    ----- -----      -----  -------
sp|Q62715|DEF2_RAT   1/1       1    94 []     1    92 []   186.0  1.1e-56
```

Figure A.2: Classification results of a query sequence using α-defensin HMM profile.

UNIVERSITY of the
WESTERN CAPE

Hydrocalculator results page



Sequence 1

- Your query sequence is:
  MRTFALLTAMLLLVALHAQAEARQARADEAAAQQQPGADDQGMAHSFTRP
  NAALPLSESARGLRCLCRRGVCQLL
  (The hydrophobic residues are coloured in blue).
- The length of the sequence is: 76 aa
- The net charge of the sequence is: +3
- The percentage of hydrophobic residues is: 50%

Mean Hydrophobicity (μ)

- The mean hydrophobicity and mean hydrophobic moment of the sequence is

Variable	CSS	Kyte & Doolittle	Eisenberg
Mean hydrophobicity	-1.26	-0.06	-0.16
Mean hydrophobic moment	0.95	0.44	0.12

Click here to see the Hydrophobic Scale used

UNIVERSITY of the WESTERN CAPE

Figure A.3: Hydrocalculator results of α-defensin sequence

SignalP results page

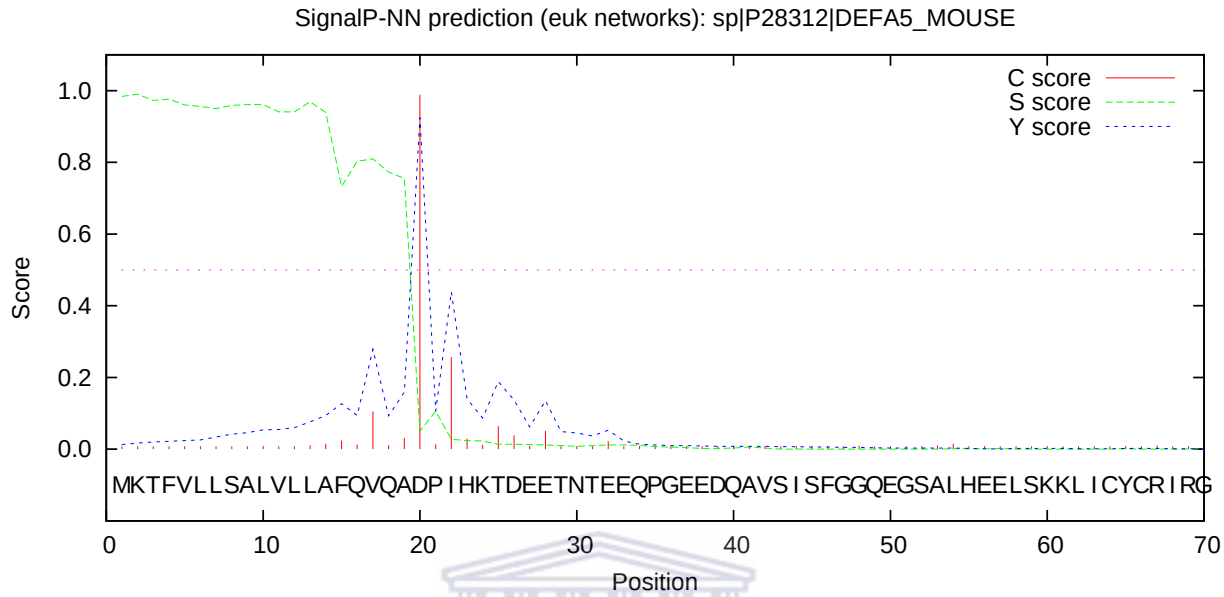


Figure A.4: SignalP results of α-defensin sequence

UNIVERSITY of the
WESTERN CAPE

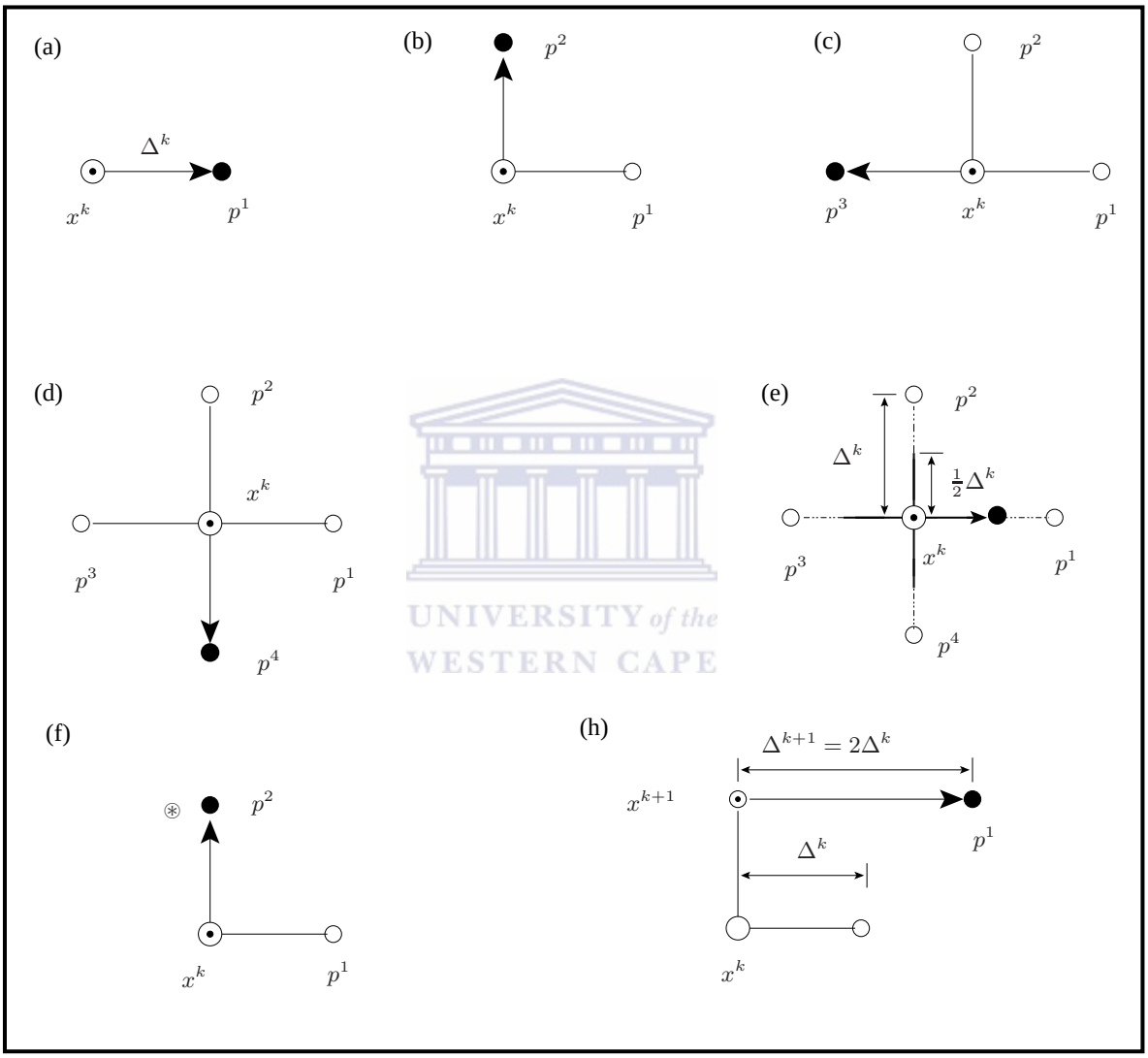
Appendix B

Supplementary material for Chapter 3

Pattern search method

This example illustrates how the previous Algorithm 1 works in $\mathbb{R}^2$. In Figure B.1, x^k is the current iterate at the k^{th} iteration and is represented by the dotted circle $\odot$. The solid circle $\bullet$ indicates the position of the trial point $p^i \in P^k$ to be examined, where $i = 1, \cdots, r$. The small open circle $\circ$ and the circled asterisk $\circledast$ represent unsuccessful and successful trial points respectively of the POLL step. The POLL step begins by evaluating the function value of the trial point $p^i \in P^k$, point by point, where $i = 1, \cdots, 4$, as shown in Figure B.1. In Figure 2.2(a), the PS method computes the trial point p^1 by a step of size Δ^k. It computes the function value at p^1. If $f(p^1) > f(x^k)$ then it examines the next trial point p^2 as shown in Figure 2.2(b). If it is not successful at p^2, i.e., $f(p^2) > f(x^k)$ then it computes p_3 as shown in Figure 2.2(c). If p^3 is still unsuccessful then the process is repeated until all the trial points in P^k are examined, i.e., until p^4 is computed as shown in Figure 2.2(d). If all the points in the POLL set P^k (i.e., p^1, p^2, p^3 and p^4) are not successful then the step size is reduced by half as shown in Figure 2.2(e), i.e., the next POLL step begins at $x^{k+1} = x^k$ with $\Delta^{k+1} = \frac{1}{2}\Delta^k$. On the other hand, suppose that the trial point p^2 is successful, i.e., $f(p^2) < f(x^k)$ as shown in Figure 2.2(f), then the whole POLL step process starts anew at $x^{k+1} = p^2$ with enlarged step size, i.e., $\Delta^{k+1} = 2\Delta^k$ as shown in Figure 2.2(h). A similar cycle as shown in (a), (b), (c) and (d) of Figure 2.2 will be repeated (if necessary) for the new POLL at x^{k+1}.

UNIVERSITY of the WESTERN CAPE

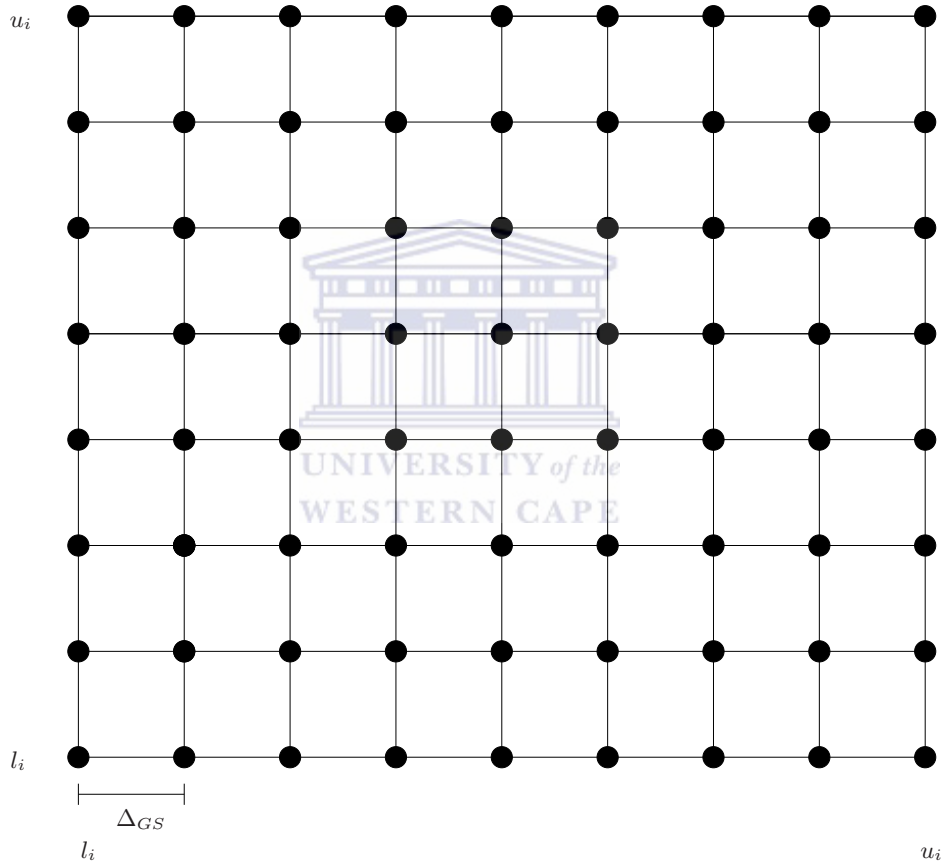


UNIVERSITY of the WESTERN CAPE

Figure B.1: Figures (a)-(h) shows how the POLL steps works in the PS method.

Grid search method

A grid search tries values of each parameter across the specified search range using geometric steps. Grid searches are computationally expensive because the model must be evaluated at many points within the grid for each parameter. For example, if a grid search is used with 20 search intervals and the svm three parameters (c, σ) then the model must be evaluated at 20×20=400 grid points. If cross-validation is used for each model evaluation, the number of actual SVM calculations would be further multiplied by the number of cross-validation folds. For large models, this approach may be computationally infeasible.



UNIVERSITY of the
WESTERN CAPE

Figure B.2: Figure shows how the Grid Search works in a two dimensional optimization problem

Negative sets

The negative set was downloaded from UniProt using the keywords below:

- actinopterygii (((golgi or cytoplasm or endoplasmic reticulum or mitochondria) NOT antimicrobial) AND length:[0 TO 100]) AND taxonomy:"Actinopterygii [7898]"
- amphibian (((golgi or cytoplasm or endoplasmic reticulum or mitochondria) NOT antimicrobial) AND length:[0 TO 100]) AND taxonomy:"Amphibia [8292]"
- arachnida (((golgi or cytoplasm or endoplasmic reticulum or mitochondria) NOT antimicrobial) AND length:[0 TO 100]) AND taxonomy:"Arachnida [6854]"
- bacteria (((golgi or cytoplasm or endoplasmic reticulum or mitochondria) NOT antimicrobial) AND length:[0 TO 100]) AND taxonomy:"Eubacterium [1730]"
- crustacea (((golgi or cytoplasm or endoplasmic reticulum or mitochondria) NOT antimicrobial) AND length:[0 TO 100]) AND taxonomy:"Crustacea [6657]"
- insecta (((golgi or cytoplasm or endoplasmic reticulum or mitochondria) NOT antimicrobial) AND length:[0 TO 100]) AND taxonomy:"Insecta [50557]"
- mammalia (((golgi or cytoplasm or endoplasmic reticulum or mitochondria) NOT antimicrobial) AND length:[0 TO 70]) AND taxonomy:"Mammalia [40674]"
- plant (((golgi or cytoplasm or endoplasmic reticulum or mitochondria) NOT antimicrobial) AND length:[0 TO 100]) AND taxonomy:"Viridiplantae [33090]"

UNIVERSITY of the WESTERN CAPE

Features indices

- **EISD840101 Consensus normalized hydrophobicity scale Kawashima et al. (2008)**

  A: 0.25, R: -1.76, N: -0.64, D: -0.72, C: 0.04, Q: -0.69, E: -0.62, G: 0.16, H: -0.40, I: 0.73, L: 0.53, K: -1.10, M: 0.26, F: 0.61, P: -0.07, S: -0.26, T: -0.18, W: 0.37, Y: 0.02, V: 0.54

- **HOPT810101 Hydrophilicity value Kawashima et al. (2008)**

  A: -0.5, R: 3.0, N: 0.2, D: 3.0, C: -1.0, Q: 0.2, E: 3.0, G: 0.0, H: -0.5, I: -1.8, L: -1.8, K: 3.0, M: -1.3, F: -2.5, P: 0.0, S: 0.3, T: -0.4, W: -3.4, Y: -2.3, V: -1.5

- **VELV850101 Electron-ion interaction potential Kawashima et al. (2008)**

  A: .03731, R: .09593, N: .00359, D: .12630, C: .08292, Q: .07606, E: .00580, G: .00499, H: .02415, I: .00000, L: .00000, K: .03710, M: .08226, F: .09460, P: .01979, S: .08292, T: .09408, W: .05481, Y: .05159, V: .00569

- **ZIMJ680101 Hydrophobicity Kawashima et al. (2008)**

  A: 0.83, R: 0.83, N: 0.09, D: 0.64, C: 1.48, Q: 0.00, E: 0.65, G: 0.10, H: 1.10, I: 3.07, L: 2.52, K: 1.60, M: 1.40, F: 2.75, P: 2.70, S: 0.14, T: 0.54, W: 0.31, Y: 2.97, V: 1.79

UNIVERSITY of the
WESTERN CAPE

- **ZIMJ680102 Bulkiness Kawashima et al. (2008)**

  A: 11.50, R: 14.28, N: 12.82, D: 11.68, C: 13.46, Q: 14.45, E: 13.57, G: 3.40, H: 13.69, I: 21.40, L: 21.40, K: 15.71, M: 16.25, F: 19.80, P: 17.43, S: 9.47, T: 15.77, W: 21.67, Y: 18.03, V: 21.57

- **ZIMJ680103 Polarity Kawashima et al. (2008)**

  A: 0.00, R: 52.00, N: 3.38, D: 49.70, C: 1.48, Q: 3.53, E: 49.90, G: 0.00, H: 51.60, I: 0.13, L: 0.13, K: 49.50, M: 1.43, F: 0.35, P: 1.58, S: 1.67, T: 1.66, W: 2.10, Y: 1.61, V: 0.13

- **JURD980101 Modified Kyte-Doolittle hydrophobicity scale Kawashima et al. (2008)**

  A: 1.10, R: -5.10, N: -3.50, D: -3.60, C: 2.50, Q: -3.68, E: -3.20, G: -0.64, H: -3.20, I: 4.50, L: 3.80, K: -4.11, M: 1.90, F: 2.80, P: -1.90, S: -0.50, T: -0.70, W: -0.46, Y: -1.3, V: 4.2

References

Abeel, T., Helleputte, T., Van de Peer, Y., Dupont, P., and Saeys, Y. (2010). Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. *Bioinformatics*, 26(3):392–8.

Abramson, M. A., Audet, C., and Dennis, J. (2004). Generalized pattern searches with derivative information. *Mathematical Programming*, 100:3–25.

Acharya, U. R., Ng, E. Y. K., Tan, J.-H., Sree, S. V., and Ng, K.-H. (2011). An integrated index for the identification of diabetic retinopathy stages using texture parameters. *J Med Syst*.

Akuffo, H., Hultmark, D., EngstÃm, A., Frohlich, D., and Kimbrell, D. (1998). Drosophila antibacterial protein, cecropin a, differentially affects non-bacterial organisms such as leishmania in a manner different from other amphipathic peptides. *Int J Mol Med*, 1(1):77–82.

Ali, M. M. and Gabere, M. N. (2010). A simulated annealing driven multi-start algorithm for bound constrained global optimization. *J. Comput. Appl. Math.*, 233(10):2661–2674.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol*, 215(3):403–10.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res*, 25(17):3389–402.

Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Birney, E., Biswas, M., Bucher, P., Cerutti, L., Corpet, F., Croning, M. D., Durbin, R., Falquet, L., Fleischmann, W., Gouzy, J., Hermjakob, H., Hulo, N., Jonassen, I., Kahn, D., Kanapin, A., Karavidopoulou, Y., Lopez, R., Marx, B., Mulder, N. J., Oinn, T. M., Pagni, M., Servant, F., Sigrist, C. J., Zdobnov, E. M., and InterPro Consortium (2000). Interpro–an integrated documentation resource for protein families, domains and functional sites. *Bioinformatics*, 16(12):1145–50.

UNIVERSITY of the
WESTERN CAPE

Audet, C., Jr., J. E. D., and Le Digabel, S. (2008). Parallel space decomposition of the mesh adaptive direct search algorithm. *SIAM J. Optim.*, 19(3):1150–1170.

Bairoch, A. and Apweiler, R. (2000). The swiss-prot protein sequence database and its supplement trembl in 2000. *Nucleic Acids Res*, 28(1):45–8.

Barutcuoglu, Z., Schapire, R. E., and Troyanskaya, O. G. (2006). Hierarchical multi-label prediction of gene function. *Bioinformatics*, 22:830–836.

Bellamy, W., Takase, M., Wakabayashi, H., Kawase, K., and Tomita, M. (1992). Antibacterial spectrum of lactoferricin b, a potent bactericidal peptide derived from the n-terminal region of bovine lactoferrin. *J Appl Bacteriol*, 73(6):472–9.

Ben-Hur, A. and Brutlag, D. (2003). Remote homology detection: a motif based approach. *Bioinformatics*, 19 Suppl 1:i26–33.

Bendtsen, J. D., Nielsen, H., von Heijne, G., and Brunak, S. (2004). Improved prediction of signal peptides: Signalp 3.0. *J Mol Biol*, 340(4):783–95.

Bjellqvist, B., Basse, B., Olsen, E., and Celis, J. E. (1994). Reference points for comparisons of two-dimensional maps of proteins from different human cell types defined in a ph scale where isoelectric points correlate with polypeptide compositions. *Electrophoresis*, 15(3-4):529–39.

Boisvert, S., Marchand, M., Laviolette, F., and Corbeil, J. (2008). Hiv-1 coreceptor usage prediction without multiple alignments: an application of string kernels. *Retrovirology*, 5:110.

Boman, H. G. (2000). Innate immunity and the normal microflora. *Immunol Rev*, 173:5–16.

Boser, B. E., Guyon, I. M., and Vapnik, N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, COLT '92, pages 144–152, New York, NY, USA. ACM.

Boulanger, N., Brun, R., Ehret-Sabatier, L., Kunz, C., and Bulet, P. (2002). Immunopeptides in the defense reactions of glossina morsitans to bacterial and trypanosoma brucei brucei infections. *Insect Biochem Mol Biol*, 32(4):369–75.

Brahmachary, M., Krishnan, S. P. T., Koh, J. L. Y., Khan, A. M., Seah, S. H., Tan, T. W., Brusic, V., and Bajic, V. B. (2004). Antimic: a database of antimicrobial sequences. *Nucleic Acids Res*, 32(Database issue):D586–9.

UNIVERSITY of the
WESTERN CAPE

Brahmachary, M., SchÃnbach, C., Yang, L., Huang, E., Tan, S. L., Chowdhary, R., Krishnan, S. P. T., Lin, C.-Y., Hume, D. A., Kai, C., Kawai, J., Carninci, P., Hayashizaki, Y., and Bajic, V. B. (2006). Computational promoter analysis of mouse, rat and human antimicrobial peptide-coding genes. *BMC Bioinformatics*, 7 Suppl 5:S8.

Breiman, L. (2001). Random forests. pages 5–32.

Brogden, K. A. (2005). Antimicrobial peptides: pore formers or metabolic inhibitors in bacteria? *Nat Rev Micro*, 3:238–250.

Bulet, P., Dimarcq, J. L., Hetru, C., Lagueux, M., Charlet, M., Hegy, G., Van Dorsselaer, A., and Hoffmann, J. A. (1993). A novel inducible antibacterial peptide of drosophila carries an o-glycosylated substitution. *J Biol Chem*, 268(20):14893–7.

Bulet, P., StÃcklin, R., and Menin, L. (2004). Anti-microbial peptides: from invertebrates to vertebrates. *Immunol Rev*, 198:169–84.

Carter, V. and Hurd, H. (2010). Choosing anti-plasmodium molecules for genetically modifying mosquitoes: focus on peptides. *Trends Parasitol*, 26(12):582–90.

Chen, W. and Luo, L. (2009). Classification of antimicrobial peptide using diversity measure with quadratic discriminant analysis. *J Microbiol Methods*, 78(1):94–6.

Cole, A. M. and Ganz, T. (2000). Human antimicrobial peptides: analysis and application. *Biotechniques*, 29(4):822–6, 828, 830–1.

Colinge, J., Masselot, A., Giron, M., Dessingy, T., and Magnin, J. (2003). Olav: towards high-throughput tandem mass spectrometry data identification. *Proteomics*, 3(8):1454–63.

Cristianini, N. and Shawe-Taylor, J. (2001). *An introduction to support vector machines and other kernel-based learning methods. Repr.* Cambridge: Cambridge University Press.

Damaševičius, R. (2010). Optimization of SVM parameters for recognition of regulatory DNA sequences. *Top*, 18(2):339–353.

de Jong, A., van Heel, A. J., Kok, J., and Kuipers, O. P. (2010). Bagel2: mining for bacteriocins in genomic data. *Nucleic Acids Res*, 38(Web Server issue):W647–51.

UNIVERSITY of the WESTERN CAPE

de Jong, A., van Hijum, S. A. F. T., Bijlsma, J. J. E., Kok, J., and Kuipers, O. P. (2006). Bagel: a web-based bacteriocin genome mining tool. *Nucleic Acids Res*, 34(Web Server issue):W273–9.

Dekkers, A. (1991). Global optimization and simulated annealing. *Mathematical Programming.*, 50:367–393.

Duan, K., Keerthi, S. S., and Poo, A. N. (2003). Evaluation of simple performance measures for tuning svm hyperparameters. *Neurocomputing*, 51:41–59.

Dutton, C. J., Haxell, M. A., McArthur, H. A. I., and Wax, R. G. (2002). *Peptide Antibiotics. Discovery, Modes of Action and Applications*. Marcel Dekker, New York, NY, USA,.

Eddy, S. R. (1998). Profile hidden markov models. *Bioinformatics*, 14(9):755–63.

Efron, B. (2004). Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *J. Am. Stat. Assoc.*, 99(465):96–104.

Ferre, R., Badosa, E., Feliu, L., Planas, M., Montesinos, E., and Bardají, E. (2006). Inhibition of plant-pathogenic bacteria by short synthetic cecropin a-melittin hybrid peptides. *Appl Environ Microbiol*, 72(5):3302–8.

Finn, R. D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J. E., Gavin, O. L., Gunasekaran, P., Ceric, G., Forslund, K., Holm, L., Sonnhammer, E. L. L., Eddy, S. R., and Bateman, A. (2010). The pfam protein families database. *Nucleic Acids Res*, 38(Database issue):D211–22.

Fjell, C. D., Hancock, R. E. W., and Cherkasov, A. (2007). Amper: a database and an automated discovery tool for antimicrobial peptides. *Bioinformatics*, 23(9):1148–55.

Gabere, M. N. (2007). Simulated annealing driven pattern search algorithms for global optimization. Master's thesis, School of Computational and Applied Mathematics, University of the Witwatersrand, Johannesburg.

Ganz, T. (2003). Defensins: antimicrobial peptides of innate immunity. *Nat Rev Immunol*, 3(9):710–20.

Garcia-Olmedo, F., Molina, A., Alamillo, J. M., and Rodriguez-Palenzuela, P. (1998). Plant defense peptides. *Biopolymers*, 47(6):479–91.

Garrido, C., Roulet, V., Chueca, N., Poveda, E., Aguilera, A., Skrabal, K., Zahonero, N., Carlos, S., GarcÃa, F., Faudon, J. L., Soriano, V., and de Mendoza, C. (2008). Evaluation of eight different bioin-

UNIVERSITY of the WESTERN CAPE

formatics tools to predict viral tropism in different human immunodeficiency virus type 1 subtypes. *J Clin Microbiol*, 46(3):887–91.

Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M. R., Appel, R. D., and Bairoch, A. (2005). *Protein Identification and Analysis Tools on the ExPASy Server*. The Proteomics Protocols Handbook, Humana Press.

Ghosh, J. K., Shaool, D., Guillaud, P., CicÃron, L., Mazier, D., Kustanovich, I., Shai, Y., and Mor, A. (1997). Selective cytotoxicity of dermaseptin s3 toward intraerythrocytic plasmodium falciparum and the underlying molecular basis. *J Biol Chem*, 272(50):31609–16.

Giangaspero, A., Sandri, L., and Tossi, A. (2001). Amphipathic alpha helical antimicrobial peptides. *Eur J Biochem*, 268(21):5589–600.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., and Bloomfield, C. D. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, pages 531–537.

Guani-Guerra, E., Santos-Mendoza, T., Lugo-Reyes, S. O., and Terán, L. M. (2010). Antimicrobial peptides: general overview and clinical implications in human health and disease. *Clin Immunol*, 135(1):1–11.

UNIVERSITY of the
WESTERN CAPE

Gueguen, Y., Garnier, J., Robert, L., Lefranc, M.-P., Mougenot, I., de Lorgeril, J., Janech, M., Gross, P. S., Warr, G. W., Cuthbertson, B., Barracco, M. A., Bulet, P., Aumelas, A., Yang, Y., Bo, D., Xiang, J., Tassanakajon, A., Piquemal, D., and Bachére, E. (2006). Penbase, the shrimp antimicrobial peptide penaeidin database: sequence-based classification and recommended nomenclature. *Dev Comp Immunol*, 30(3):283–8.

Hammami, R., Ben Hamida, J., Vergoten, G., and Fliss, I. (2009). Phytamp: a database dedicated to antimicrobial plant peptides. *Nucleic Acids Res*, 37(Database issue):D963–8.

Hammami, R., Zouhir, A., Ben Hamida, J., and Fliss, I. (2007). Bactibase: a new web-accessible database for bacteriocin characterization. *BMC Microbiol*, 7:89.

Hancock, R. E. and Diamond, G. (2000). The role of cationic antimicrobial peptides in innate host defences. *Trends Microbiol*, 8(9):402–10.

Hancock, R. E. and Lehrer, R. (1998). Cationic peptides: a new source of antibiotics. *Trends Biotechnol*, 16(2):82–8.

Hancock, R. E. W. and Chapple, D. S. (1999). Peptide antibiotics. *Antimicrobial Agents Chemotherapy.*, 43(6):1317–1323.

Harris, D. J. (2003). Can you bank on genbank? *Trends in Ecology & Evolution*, 18(7):317–319.

Hedar, A. and Fukushima, M. (2004). Heuristic pattern search and its hybridization with simulated annealing for nonlinear global optimization. *Optim. Methods Softw.*, 19(3-4):291–308.

Hoffmann, J. A. and Hetru, C. (1992). Insect defensins: inducible antibacterial peptides. *Immunol Today*, 13(10):411–5.

Howe, D., Costanzo, M., Fey, P., Gojobori, T., Hannick, L., Hide, W., Hill, D. P., Kania, R., Schaeffer, M., St Pierre, S., Twigger, S., White, O., and Rhee, S. Y. (2008). Big data: The future of biocuration. *Nature*, 455(7209):47–50.

Hu, Y. and Aksoy, S. (2005). An antimicrobial peptide with trypanocidal activity characterized from glossina morsitans morsitans. *Insect Biochem Mol Biol*, 35(2):105–15.

UNIVERSITY of the
WESTERN CAPE

Huang, J. and Ling, C. X. (2007). Constructing new and better evaluation measures for machine learning.

Jaakkola, T., Diekhans, M., and Haussler, D. (2000). A discriminative framework for detecting remote protein homologies. *J Comput Biol*, 7(1-2):95–114.

Jin, B., Muller, B., Zhai, C., and Lu, X. (2008). Multi-label literature classification based on the gene ontology graph. *BMC Bioinformatics*, 9:1–15. 10.1186/1471-2105-9-525.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. *Machine Learning: ECML-98*, 1398:137–142.

Joachims, T. (1999). Making large-scale svm learning practical. In SchÃlkopf, B., Burges, C., and Smola, A., editors, *Advances in Kernel Methods - Support Vector Learning*. MIT Press.

Käll, L., Storey, J. D., MacCoss, M. J., and Noble, W. S. (2008a). Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. *J Proteome Res*, 7(1):29–34.

Käll, L., Storey, J. D., MacCoss, M. J., and Noble, W. S. (2008b). Posterior error probabilities and false discovery rates: two sides of the same coin. *J Proteome Res*, 7(1):40–4.

Käll, L., Storey, J. D., and Noble, W. S. (2008c). Non-parametric estimation of posterior error probabilities associated with peptides identified by tandem mass spectrometry. *Bioinformatics*, 24(16):i42–8.

Käll, L., Storey, J. D., and Noble, W. S. (2009). Qvality: non-parametric estimation of q-values and posterior error probabilities. *Bioinformatics*, 25(7):964–6.

Kamysz, W., Okrüj, M., and Łukasiak, J. (2003). Novel properties of antimicrobial peptides. *Acta Biochim Pol*, 50(2):461–9.

Kapetanovic, I. M., Rosenfeld, S., and Izmirlian, G. (2004). Overview of commonly used bioinformatics methods and their applications. *Ann N Y Acad Sci*, 1020:10–21.

Kawashima, S., Pokarowski, P., Pokarowska, M., Kolinski, A., Katayama, T., and Kanehisa, M. (2008). Aaindex: amino acid index database, progress report 2008. *Nucleic Acids Res*, 36(Database issue):D202–5.

Klammer, A. A. and MacCoss, M. J. (2006). Effects of modified digestion schemes on the identification of proteins from complex mixtures. *J Proteome Res*, 5(3):695–700.

Kolda, T. G., Lewis, R. M., and Torczon, V. (2003). Optimization by direct search: New perspectives on some classical and modern methods. *SIAM Review*, 45(3):385–482.

UNIVERSITY of the
WESTERN CAPE

Korn, F., Sidiropoulos, N., Faloutsos, C., Siegel, E., and Protopapas, Z. (2007). Fast nearest neighbor search in medical image databases. pages 215–226.

Kuang, R., Ie, E., Wang, K., Wang, K., Siddiqi, M., Freund, Y., and Leslie, C. (2005). Profile-based string kernels for remote homology detection and motif extraction. *J Bioinform Comput Biol*, 3(3):527–50.

Kyrpides, N. C. (1999). Genomes online database (gold 1.0): a monitor of complete and ongoing genome projects world-wide. *Bioinformatics*, 15(9):773–4.

Lata, S., Mishra, N. K., and Raghava, G. P. S. (2010). Antibp2: improved version of antibacterial peptide prediction. *BMC Bioinformatics*, 11 Suppl 1:S19.

Lata, S., Sharma, B. K., and Raghava, G. P. S. (2007). Analysis and prediction of antibacterial peptides. *BMC Bioinformatics*, 8:263.

Lawson, D., Arensburger, P., Atkinson, P., Besansky, N. J., Bruggner, R. V., Butler, R., Campbell, K. S., Christophides, G. K., Christley, S., Dialynas, E., Hammond, M., Hill, C. A., Konopinski, N., Lobo,

N. F., MacCallum, R. M., Madey, G., Megy, K., Meyer, J., Redmond, S., Severson, D. W., Stinson, E. O., Topalis, P., Birney, E., Gelbart, W. M., Kafatos, F. C., Louis, C., and Collins, F. H. (2009). Vectorbase: a data resource for invertebrate vector genomics. *Nucleic Acids Res*, 37(Database issue):D583–7.

Lee, S. Y., Kim, S., Kim, S. S., Cha, S. J., Kwon, Y. K., Moon, B. R., and Lee, B. J. (2004). Application of decision tree for the classification of antimicrobial peptide. *Genomics and Informatics.*, 2(3):121–125.

Lehrer, R. I. and Ganz, T. (2002). Defensins of vertebrate animals. *Curr Opin Immunol*, 14(1):96–102.

Leslie, C., Eskin, E., and Noble, W. S. (2002). The spectrum kernel: a string kernel for svm protein classification. *Pac Symp Biocomput*, pages 564–75.

Leslie, C. S., Eskin, E., Cohen, A., Weston, J., and Noble, W. S. (2004). Mismatch string kernels for discriminative protein classification. *Bioinformatics*, 20(4):467–76.

Letunic, I., Doerks, T., and Bork, P. (2009). Smart 6: recent updates and new developments. *Nucleic Acids Res*, 37(Database issue):D229–32.

Li, W. and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658–9.

Li, W., Jaroszewski, L., and Godzik, A. (2002). Tolerating some redundancy significantly speeds up clustering of large protein databases. *Bioinformatics*, 18(1):77–82.

Liao, L. and Noble, W. S. (2002). Combining pairwise sequence similarity and support vector machines for remote protein homology detection. pages 225–232.

Lin, S., Lee, Z., Chen, S., and Tseng, T. (2008). Parameter determination of support vector machine and feature selection using simulated annealing approach. *Applied Soft Computing*, 8(4):1505–1512. Soft Computing for Dynamic Data Mining.

Matsuzaki, K. (1999). Why and how are peptide-lipid interactions utilized for self-defense? magainins and tachyplesins as archetypes. *Biochim Biophys Acta*, 1462(1-2):1–10.

Maxwell, A. I., Morrison, G. M., and Dorin, J. R. (2003). Rapid sequence divergence in mammalian beta-defensins by adaptive evolution. *Mol Immunol*, 40(7):413–21.

Mika, S. and Rost, B. (2004). Protein names precisely peeled off free text. *Bioinformatics*, 20 Suppl 1:i241–7.

UNIVERSITY of the WESTERN CAPE

Momma, M. and Bennett, K. P. (2002). A pattern search method for model selection of support vector regression.

Moore, R. E., Young, M. K., and Lee, T. D. (2002). Qscore: an algorithm for evaluating sequest database search results. *J Am Soc Mass Spectrom*, 13(4):378–86.

Mulvenna, J. P., Wang, C., and Craik, D. J. (2006). Cybase: a database of cyclic protein sequence and structure. *Nucleic Acids Res*, 34(Database issue):D192–4.

Nagarajan, V., Kaushik, N., Murali, B., Zhang, C., Lakhera, S., Elasri, M. O., and Deng, Y. (2006). A fourier transformation based method to mine peptide space for antimicrobial activity. *BMC Bioinformatics*, 7 Suppl 2:S2.

Noble, W. S. (2006). What is a support vector machine? *Nat Biotechnol*, 24(12):1565–7.

Noble, W. S. (2009). How does multiple testing correction work? *Nat Biotechnol*, 27(12):1135–7.

Pan, W. (2002). A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics*, 18(4):546–54.

Park, I. Y., Park, C. B., Kim, M. S., and Kim, S. C. (1998). Parasin i, an antimicrobial peptide derived from histone h2a in the catfish, parasilurus asotus. *FEBS Lett*, 437(3):258–62.

Passerini, A., Punta, M., Ceroni, A., Rost, B., and Frasconi, P. (2006). Identifying cysteines and histidines in transition-metal-binding sites using support vector machines and neural networks. *Proteins*, 65(2):305–16.

Pavlidis, P., Cai, J., Weston, J., and Grundy, W. N. Wn: Gene functional classification from heterogeneous data.

Perriére, G. and Gouy, M. (1996). Www-query: an on-line retrieval system for biological sequence banks. *Biochimie*, 78(5):364–9.

Pieper, U., Eswar, N., Webb, B. M., Eramian, D., Kelly, L., Barkan, D. T., Carter, H., Mankoo, P., Karchin, R., Marti-Renom, M. A., Davis, F. P., and Sali, A. (2009). Modbase, a database of annotated comparative protein structure models and associated resources. *Nucleic Acids Res*, 37(Database issue):D347–54.

Popescul, A., Popescul, R., and Ungar, L. H. (2003). Structural logistic regression for link analysis.

Pounds, S. and Morris, S. W. (2003). Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values. *Bioinformatics*, 19(10):1236–42.

Prosperi, M. C. F., Fanti, I., Ulivi, G., Micarelli, A., De Luca, A., and Zazzi, M. (2009). Robust supervised and unsupervised statistical learning for hiv type 1 coreceptor usage analysis. *AIDS Res Hum Retroviruses*, 25(3):305–14.

Rinaldi, A. C. (2002). Antimicrobial peptides from amphibian skin: an expanding scenario. *Curr Opin Chem Biol*, 6(6):799–804.

Samakovlis, C., Kimbrell, D. A., Kylsten, P., EngstrÃm, A., and Hultmark, D. (1990). The immune response in drosophila: pattern of cecropin expression and biological activity. *EMBO J*, 9(9):2969–76.

Samanta, B., Al-Balushi, K. R., and Al-Araimi, S. A. (2006). Artificial neural networks and genetic algorithm for bearing fault detection. *Soft Comput*, 10(3):264–271.

Sander, O., Sing, T., Sommer, I., Low, A. J., Cheung, P. K., Harrigan, P. R., Lengauer, T., and Domingues, F. S. (2007). Structural descriptors of gp120 v3 loop for the prediction of hiv-1 coreceptor usage. *PLoS Comput Biol*, 3(3):e58.

Sasson, O., Vaaknin, A., Fleischer, H., Portugaly, E., Bilu, Y., Linial, N., and Linial, M. (2003). Protonet: hierarchical classification of the protein space. *Nucleic Acids Res*, pages 348–352.

Schietgat, L., Vens, C., Struyf, J., Blockeel, H., Kocev, D., and Dzeroski, S. (2010). Predicting gene function using hierarchical multi-label decision tree ensembles. *BMC Bioinformatics*, 11(1):2.

Schnoes, A. M., Brown, S. D., Dodevski, I., and Babbitt, P. C. (2009). Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS Comput Biol*, 5(12):e1000605.

Seebah, S., Suresh, A., Zhuo, S., Choong, Y. H., Chua, H., Chuon, D., Beuerman, R., and Verma, C. (2007). Defensins knowledgebase: a manually curated database and information source focused on the defensins family of antimicrobial peptides. *Nucleic Acids Res*, 35(Database issue):D265–8.

Shai, Y. (2002). Mode of action of membrane active antimicrobial peptides. *Biopolymers*, 66(4):236–48.

Sigrist, C. J. A., Cerutti, L., de Castro, E., Langendijk-Genevaux, P. S., Bulliard, V., Bairoch, A., and Hulo, N. (2010). Prosite, a protein domain database for functional characterization and annotation. *Nucleic Acids Res*, 38(Database issue):D161–6.

UNIVERSITY of the
WESTERN CAPE

Simmaco, M., Mignogna, G., and Barra, D. (1998). Antimicrobial peptides from amphibian skin: what do they tell us? *Biopolymers*, 47(6):435–50.

Simpson, P. K. (1990). *Artificial Neural Systems*. Pergamon Press.

Skrabal, K., Low, A. J., Dong, W., Sing, T., Cheung, P. K., Mammano, F., and Harrigan, P. R. (2007). Determining human immunodeficiency virus coreceptor use in a clinical setting: degree of correlation between two phenotypic assays and a bioinformatic model. *J Clin Microbiol*, 45(2):279–84.

Soong, T.-T., Wrzeszczynski, K. O., and Rost, B. (2008). Physical protein-protein interactions predicted from microarrays. *Bioinformatics*, 24(22):2608–14.

Statnikov, A., Aliferis, C. F., Tsamardinos, I., Hardin, D., and Levy, S. (2005). A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. *Bioinformatics*, 21(5):631–43.

Steiner, H., Hultmark, D., Engström, Å., Bennich, H., and Boman, H. G. (1981). Sequence and specificity of two antibacterial proteins involved in insect immunity. *Nature*, 292(5820):246–8.

Storey, J. D. and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A*, 100(16):9440–5.

UNIVERSITY of the
WESTERN CAPE

Strimmer, K. (2008a). fdrtool: a versatile r package for estimating local and tail area-based false discovery rates. *Bioinformatics*, 24(12):1461–2.

Strimmer, K. (2008b). A unified approach to false discovery rate estimation. *BMC Bioinformatics*, 9:303.

Thomas, S., Karnik, S., Barai, R. S., Jayaraman, V. K., and Idicula-Thomas, S. (2010). Camp: a useful resource for research on antimicrobial peptides. *Nucleic Acids Res*, 38(Database issue):D774–80.

Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994). Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*, 22(22):4673–80.

Tipping, M. E. and Bishop, C. M. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society, Series B*, 61:611–622.

Torczon, V. (1991). On the convergence of the multidirectional search algorithm.

Tossi, A., Sandri, L., and Giangaspero, A. (2002). New consensus hydrophobicity scale extended to non-proteinogenic amino acids. *Peptide*, pages 416–417.

Valanne, S., Wang, J.-H., and RÃmet, M. (2011). The drosophila toll signaling pathway. *J Immunol*, 186(2):649–56.

van 't Hof, W., Veerman, E. C., Helmerhorst, E. J., and Amerongen, A. V. (2001). Antimicrobial peptides: properties and applicability. *Biol Chem*, 382(4):597–619.

Vapnik, V. N. (2000). *The Nature of Statistical Learning Theory*. Springer-Verlag, New York.

Verma, C., Seebah, S., Low, S. M., Zhou, L., Liu, S. P., Li, J., and Beuerman, R. W. (2007). Defensins: antimicrobial peptides for therapeutic development. *Biotechnol J*, 2(11):1353–9.

Vizioli, J. and Salzet, M. (2002). Antimicrobial peptides from animals: focus on invertebrates. *Trends Pharmacol Sci*, 23(11):494–6.

Wade, D. and Englund, J. (2002). Synthetic antibiotic peptides database. *Protein Pept Lett*, 9(1):53–7.

Wang, C. K. L., Kaas, Q., Chiche, L., and Craik, D. J. (2008a). Cybase: a database of cyclic protein sequences and structures, with applications in protein discovery and engineering. *Nucleic Acids Res*, 36(Database issue):D206–10.

UNIVERSITY of the
WESTERN CAPE

Wang, G., Li, X., and Wang, Z. (2009). Apd2: the updated antimicrobial peptide database and its application in peptide design. *Nucleic Acids Res*, 37(Database issue):D933–7.

Wang, J., Hu, C., Wu, Y., Stuart, A., Amemiya, C., Berriman, M., Toyoda, A., Hattori, M., and Aksoy, S. (2008b). Characterization of the antimicrobial peptide attacin loci from glossina morsitans. *Insect Mol Biol*, 17(3):293–302.

Wang, Z. and Wang, G. (2004). : the antimicrobial peptide database. *Nucleic Acids Res*, 32(Database issue):D590–2.

Wasserman, W. W. and Fickett, J. W. (1998). Identification of regulatory regions which confer muscle-specific gene expression. *J Mol Biol*, 278(1):167–81.

Weston, J., Leslie, C., Ie, E., Zhou, D., Elisseeff, A., and Noble, W. S. (2005). Semi-supervised protein classification using cluster kernels. *Bioinformatics*, 21(15):3241–7.

Whitmore, L., Chugh, J. K., Snook, C. F., and Wallace, B. A. (2003). The peptaibol database: a sequence and structure resource. *J Pept Sci*, 9(11-12):663–5.

Wilkins, M. R., Lindskog, I., Gasteiger, E., Bairoch, A., Sanchez, J. C., Hochstrasser, D. F., and Appel, R. D. (1997). Detailed peptide characterization using peptidemass–a world-wide-web-accessible tool. *Electrophoresis*, 18(3-4):403–8.

Wu, Y., Wei, B., Liu, H., Li, T., and Rayner, S. (2011). Mirpara: a svm-based software tool for prediction of most probable microrna coding regions in genome scale sequences. *BMC Bioinformatics*, 12(1):107.

Yang, Z. R. (2004). Biological applications of support vector machines. *Brief Bioinform*, 5(4):328–38.

Yassine, H. and Osta, M. A. (2010). Anopheles gambiae innate immunity. *Cell Microbiol*, 12(1):1–9.

Ye, J., Member, S., Li, Q., and Member, S. (2005). A two-stage linear discriminant analysis via qr-decomposition. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 27:929–941.

Yeaman, M. R. and Yount, N. Y. (2003). Mechanisms of antimicrobial peptide action and resistance. *Pharmacol Rev*, 55(1):27–55.

Yount, N. Y., Bayer, A. S., Xiong, Y. Q., and Yeaman, M. R. (2006). Advances in antimicrobial peptide immunobiology. *Biopolymers*, 84(5):435–58.

Zasloff, M. (2002). Antimicrobial peptides of multicellular organisms. *Nature*, 415(6870):389–95.

Zernov, V. V., Balakin, K. V., Ivaschenko, A. A., Savchuk, N. P., and Pletnev, I. V. (2003). Drug discovery using support vector machines. the case studies of drug-likeness, agrochemical-likeness, and enzyme inhibition predictions. *J Chem Inf Comput Sci*, 43(6):2048–56.

Zhang, H., Ling, C. X., and Zhao, Z. (2005). The learnability of naive bayes. In *In: Proceedings of Canadian Artificial Intelligence Conference*, pages 432–441. AAAI Press.

Zhao, H. W., Zhou, D., and Haddad, G. G. (2011). Antimicrobial peptides increase tolerance to oxidant stress in drosophila melanogaster. *J Biol Chem*, 286(8):6211–8.

Zheng, L., Li, X., Li, F., Yan, X., Wang, Y., and Wang, Z. (2011). Automatic classification of lip color based on svm in traditional chinese medicine inspection. *Image and signal processing (CISP),2010 3rd International Congress*, 28(1):7–11.

UNIVERSITY of the
WESTERN CAPE