Highly Accurate Real-Time Classification of Channel-Captured DNA Termini

Stephen Winters-Hilt

Center for Biomolecular Science & Engineering, University of California, Santa Cruz
Computer Science Department, University of California, Santa Cruz

Abstract. A computational method is briefly described for classification of individual DNA molecules measured by an α-hemolysin channel detector. Classification is performed with better than 99% accuracy for DNA hairpin molecules that differ only in their terminal Watson-Crick base pairs. Signal classification was initially done on synthetic data streams, where sampling on real mixtures of hairpins was modeled in order to establish performance metrics (i.e., where train and test data were of known type, via single-species data files). Signal classification was then performed on observations from real mixtures of DNA hairpins. Hidden Markov Models (HMMs) were used with Expectation/Maximization for de-noising and for associating a feature vector with the ionic current blockade of the DNA molecule. Support Vector Machines (SVMs) were used as discriminators, and were the focus of off-line training. A multi-class SVM architecture was designed to place less discriminatory load on weaker discriminators, and novel SVM kernels were used to boost discrimination strength. The tuning on HMMs and SVMs enabled biophysical analysis of the captured molecule states and state-transitions; structure revealed in the biophysical analysis was used for better feature selection. This analysis is presented in more detail, but with less discussion, in Winters-Hilt et al. 2003.

INTRODUCTION

Molecular classification using nanometer-scale channels provides a powerful new tool for efforts in biophysics and biotechnology. Such nanopore detectors use a nanometer-scale channel to relate ionic current blockade measurements to single molecule translocation by the pore (Akeson et al., 1999; Kasianowicz et al., 1996; Meller et al., 2000; Meller et al. 2001), or to single molecule capture by the pore (Vercoutere et al., 2001). Biologically based α-hemolysin channels are elegant in this regard in that they self-assemble in lipid bilayers (Gouaux et al., 1994; Song et al., 1996), thereby providing inexpensive and reproducible nanopores. The size of the α-hemolysin pore is also well suited to DNA measurement in that single-stranded DNA (ssDNA) translocates while double stranded DNA (dsDNA) does not, being held instead in a vestibule of the pore (Vercoutere et al., 2001). For end-capture on dsDNA, extensive characterization of ionic current blockades is possible.

Modifications to the α-hemolysin channel have been examined (Bayley 2000), and semiconductor nanopores are being developed (Li et al., 2001). In this paper a brief description is given of how a nanopore detector, coupled with pattern recognition, can be used to discriminate between DNA hairpin termini with high accuracy. A more extensive discussion can be found in (Winters-Hilt et al., 2003).

In the nanopore signal analysis a Hidden Markov Model (HMM) is used to extract a feature vector from each blockade example. HMMs (Chung et al., 1990; Chung and Gage, 1998; Colquhoun and Sigworth, 1995) can characterize current blockades by identifying a sequence of sub-blockades as a sequence of state emissions. The parameters of an HMM are usually estimated using a method called Expectation/Maximization (Durbin 1998). Although HMMs can be used to discriminate among several classes of input, multi-class computational scalability tends to favor their use as feature extractors. In particular, HMMs are well suited to extraction of aperiodic information embedded in stochastic sequential data. Support Vector Machines (SVMs) are then used to classify the feature vectors obtained by the HMM (for each individual blockade event). SVMs are fast, easily trained, discriminators (Vapnik 1998; Burges 1998), for which strong discrimination is possible without the over-fitting complications common to neural net discriminators (Vapnik 1998).

METHODS

Nanopore Implementation And DNA Hairpin Design

Each experiment was conducted using one α-hemolysin channel inserted into a diphytanoyl-phosphatidylcholine/hexadecane bilayer, where the bilayer was formed across a 20-micron diameter horizontal Teflon aperture (Vercoutere et al., 2001). The bilayer separates two seventy-microliter chambers containing 1.0 M KCl buffered at pH 8.0 (10 mM HEPES/KOH). The nine base-pair hairpin molecules examined share an eight base-pair hairpin core sequence, with addition of one of the four permutations of Watson-Crick base-pairs that may exist at the blunt end terminus, i.e., 5'-G•C-3', 5'-C•G-3', 5'-T•A-3', and 5'-A•T-3'. Denoted 9GC, 9CG, 9TA, and 9AT, respectively. The full sequence for the 9CG hairpin is 5' CTTCGAACGTTTTCGTTCGAAG 3', where the base-pairing region is underlined. An eight base-pair DNA hairpin with a 5'-G•C-3' terminus was also tested. The prediction that each hairpin would adopt one base-paired structure was tested and confirmed using the DNA mfold server (http://bioinfo.math.rpi.edu/~mfold/dna/form1.cgi), which is based in part on data from (SantaLucia 1998). The nanopore construction and the DNA synthesis tools are described in (Winters-Hilt et al., 2003).

Sampling Protocol And Signal Acquisition

The solution sampling protocol used periodic reversal of the applied potential to accomplish the capture and ejection of single DNA molecules (added to the cis chamber in 20 μM concentrations). The current blockade data was filtered at 10 kHz bandwidth using an analog low pass Bessel filter and recorded at 20 μs intervals using an Axopatch 200B amplifier coupled to an Axon Digidata 1200 digitizer (Axon Instruments, Foster City, CA). A time-domain finite state automaton (FSA; Cormen et al. 1989) with eight states performed the identification and acquisition on the first 100 ms of blockade signal (Acquisition Stage, Figure 1). The effective duty cycle for acquiring the desired 100 ms blockade measurements was one reading every 0.4 seconds. Further details on voltage toggling protocol the time-domain FSA are in (Winters-Hilt et al., 2003).

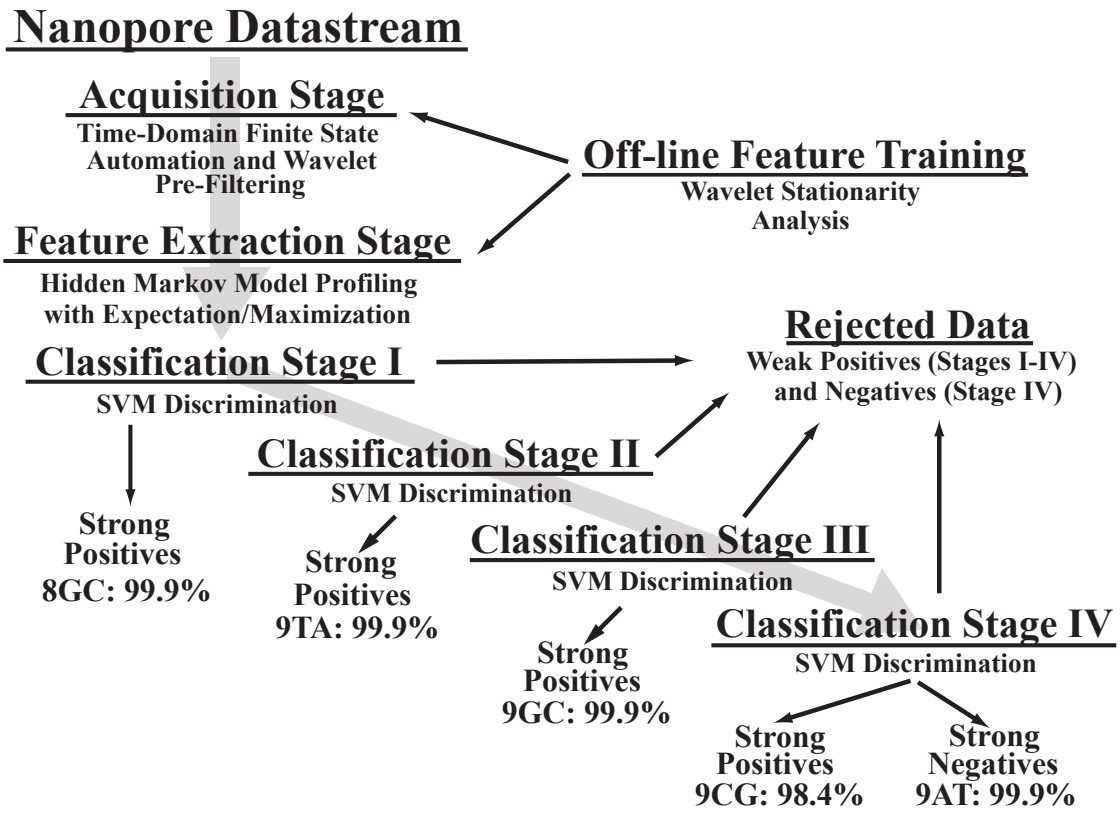

FIGURE 1. Signal acquisition was performed using a time-domain, thresholding, Finite State Automaton. This was followed by adaptive pre-filtering using a wavelet-domain Finite State Automaton. Feature extraction on those acquired channel blockades was done by Hidden Markov Model processing; and classification was done by Support Vector Machine. The optimal SVM architecture is shown for classification of molecules 9CG, 9GC, 9TA, 9AT, and 8GC. The linear tree multi-class SVM architecture benefits from strong signal skimming and weak signal rejection along the line of decision nodes. Scalability to larger multi-class problems is possible since the main on-line computational cost is at the Hidden Markov Model feature extraction stage. The accuracy shown is for single-species mixture identification upon completing the 15^{th} single molecule sampling/classification (in approx. 6 seconds).

Signal Preprocessing And Feature Extraction

Each 100 ms signal acquired by the time-domain FSA consisted of a sequence of 5000 sub-blockade levels (with the 20 μs analog-to-digital sampling). Signal preprocessing was then used for adaptive low-pass filtering. For the data sets examined the preprocessing led to length compression on the sample sequence from 5000 to 625 samples (later HMM processing then only required construction of a dynamic programming table with 625 columns). The signal preprocessing makes use of an off-line wavelet stationarity analysis (Off-line Wavelet Stationarity Analysis, Figure 1, Diserbo et al., 2000). With completion of preprocessing, an HMM (Durbin 1998) was used to remove noise from the acquired signals, and to extract features from them (Feature Extraction Stage, Fig. 1). The HMM was implemented with fifty states, corresponding to current blockades in 1% increments ranging from 20% residual current to 69% residual current. The HMM states, numbered 0 to 49, corresponded to the 50 different current blockade levels in the discrete sequences that it processed. The state emission parameters of the HMM were initially set so that the state j, 0 <= j <= 49 corresponding to level L = j+20, could emit all possible levels, with the probability distribution over emitted levels set to a discretized Gaussian with mean L and unit variance. All transitions between states were possible, and initially were equally likely. Each blockade signature was de-noised by 5 rounds of Expectation-Maximization (EM) training on the parameters of the HMM. After the EM iterations, 150 parameters were extracted from the HMM (further details in Winters-Hilt et al., 2003). The resulting parameter vector, normalized such that vector components sum to unity, was used to represent the acquired signal in discrimination at the Support Vector Machine stages.

Classification Training

The normalized feature vectors obtained from the feature extraction stage were classified using binary Support Vector Machines (SVMs). Binary SVMs are based on a decision-hyperplane heuristic that incorporates structural risk management by attempting to obtain a training-instance void, or "margin," around the decision hyperplane. Binary SVMs can be grouped into a classifier tree to perform multi-class discrimination, and this was done here for the five classes of DNA hairpin (shown in classification stages I-IV in Figure 1). Tuning on the multi-class SVM architecture itself was done for performance optimization, and separate tuning was also done on the polarization strength used in the data cleaning. Tuning was also done on the SVM internals, over families of kernels based on regularized distances (Jaakkola 1998) and regularized information divergences. In the former case, the squared Euclidean distance between feature vectors $\mathbf{x}$ and $\mathbf{y}$, $\mathbf{d^2(x,y) = \Sigma_k(x_k-y_k)^2}$, also known as the squared $\mathbf{l_2}$-norm on $\mathbf{(x-y)}$, $\mathbf{[l_2(x-y)]^2 = d^2(x,y)}$, is associated with the Gaussian kernel: $\mathbf{K_G(x,y) = exp(-d^2(x,y)/2\sigma^2)}$. In the latter case, the information divergence (relative entropy) between probability vectors $\mathbf{x}$ and $\mathbf{y}$, $\mathbf{D(x||y) = \Sigma_k x_k log(x_k/y_k)}$, can be associated with the "Entropic kernel:" $\mathbf{K_E(x,y) = exp(-[D(x||y)+D(y||x)]/2\sigma^2)}$. The

terminating SVM node of the classifier tree (stage IV in Figure 1) used the Entropic kernel. The other nodes of the classifier tree used a regularized-distance type kernel, the "Indicator kernel," based on the square root of the $\mathbf{l_1}$-norm, where $\mathbf{l_1(x-y) = \Sigma_k|x_k-y_k|}$, with kernel $\mathbf{K_I(x,y) = exp(-\sqrt{l_1(x-y)}/2\sigma^2)}$. The kernels considered were not restricted by Mercer's conditions. Instead, attention was focused on exploring kernels based on regularized information divergences as a parallel to the very successful kernels based on regularized distances (such as the Gaussian kernel). The Gaussian kernel (which satisfies Mercer's conditions) was outperformed in all cases studied by the Entropic and Indicator kernels.

Discriminator Implementation

The SVM discriminators were trained by solving their KKT relations using the Sequential Minimal Optimization (SMO) procedure (Platt 1998). A Chunking (Osuna et al., 1997; Joachims, 1998) variant of SMO was employed to manage the large training task at each SVM node. The multi-class SVM training was based on over ten thousand blockade signatures for each DNA hairpin species. The data cleaning needed on the training data was accomplished by an extra SVM training round (further details on data cleaning in Winters-Hilt et al., 2003).

Testing Protocol

The test data consisted of over two thousand blockade signals for each DNA hairpin species and was drawn from experiments that were run on days (and nanopores) different from those used to acquire the training data. Testing on single-species mixture calling was done directly, with classification on observations from single-species solutions in the cis chamber. One goal of the study was to find how many classification attempts were required to classify the single-species solutions with very high confidence. Scoring was possible by tracking the known labels on the test data. For the mixture tests some of the train data was used for an added calibration. An extra calibration was required because true mixtures of hairpins are sensitive to the different (entropic) acceptance rates and (discriminator) rejection rates by the nanopore instrument for the different hairpin species.

RESULTS

We were able to determine which of five species of DNA hairpin had been added to the cis chamber of the nanopore device. This was done in less than six seconds with 99.6% accuracy. The five DNA hairpins consisted of four hairpins that differed only in their terminal base-pairs, and a control hairpin. These results were data drawn from nanopores established on days other than those used to generate the training data. At 75% weak signal rejection, approximately 15 classification attempts were needed to classify the type of single-species solution being sampled; final solution classification was obtained in six seconds on average. If training and testing were done on data drawn from the same set of days of nanopore operation, albeit different samples,

99.9% calling was obtained with 15% rejection, and throughput was about one call every half second.

Identification of two hairpins in mixtures was also attempted. Figure 2 shows the percentage of 9TA classification in a 3:1 mixture of 9TA to 9GC. (Although the mixture preparations are estimated to be ±10% of their stated mixture ratios, calibration and testing of aliquots from the same mixture compensates for such common error.) The assay on 9TA concentration asymptotes to 75% ± 1%, consistent with the 3:1 ratio, and the assay error drops to 1% after approximately 100 individual molecule classification attempts (completed in 40 seconds).

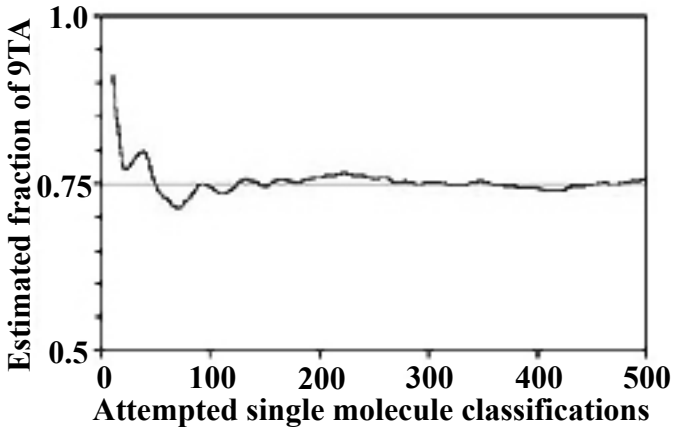

FIGURE 2. Classification on a 3:1 mixture of 9TA and 9GC hairpin molecules as a function of single molecule acquisitions.

HMM/EM characterization on the five classes of hairpin signatures revealed the existence of two major conductance blockade levels, one minor level intermediate between them, and one to three other statistically relevant levels depending on the hairpin. By examining the transition probabilities between the various levels it was found that blockades typically began in the less common intermediate level and from there almost always transitioned to the greater conductance blockade level.

DISCUSSION

Calibration And Feature Extraction By HMM

A single HMM/EM process was used to perform feature extractions in the experiments. If separate HMMs were used to model each species, the HMM/EM

processing could also be operated in a discriminative mode. This requires multiple HMM/EM evaluations (one for each species) on each unknown signal as it is observed. Increased computational burden would thus be added at the worst place: the expensive feature extraction stage. For future work, semi-scalable, species-specific processing is being considered for the HMM/EM in an indirect manner, by using prior HMM/EM characterization of the species to identify a reduced set of features relevant to each species. Traditional signal analysis on the data, using power spectral analysis methods, reveals approximately Lorentzian noise, indicating an approximately two-state random process from which rate constants for transitions back and forth (or time constants) can be extracted. In comparison, the HMM/EM reduced feature set is far more detailed, and can provide a complete (statistically optimized) physical characterization of the captured molecule, via blockade states, their time-constants, and allowed state transitions.

Tests with mixtures of hairpins required an added calibration due to the nanopore's different acceptance rates for different hairpins (i.e., there are different free energy barriers to capture). This finding was consistent with a model for hairpin capture in which hairpins are captured by an entropically accessible binding site. It is also in agreement with the brief intermediate level state typically observed at the start of the signal blockades.

Classification By SVM Hierarchy

Novel SVM kernels were used to obtain the results described here. The novel kernels are based on a generalization from regularized square-distances to regularized information divergences. One of the kernels (the Entropic kernel at Classification Stage IV in Figure 1) used the Kullback-Leibler information divergence (Cover and Thomas, 1991). Entropic-type kernels may offer advantages when all or part of the feature vector can be interpreted as a probability vector. SVM confidence for a classification is a function of the distance from the feature vector coordinates (in feature space) to the separating, discriminatory, hyperplane (where greater distance represents higher confidence in discriminating between two signals). Since the kernel defines the notion of distance, tuning on kernels also provides a powerful means to manage rejection behavior. Multi-class SVM discrimination can then obtained by grouping binary SVMs into a decision-tree architecture (Vapnik, 1998; Bredensteiner and Bennett, 1999) using rejection of low confidence data at earlier stages to postpone decisions to more appropriate later stages.

Re-establishing the α-hemolysin channel on a day-to-day basis presents a major complication to the pattern recognition task. The class training data that would normally map to a single cluster is shattered into a cluster of clusters, with greater dispersion and class overlap in the SVM feature vector space. SVM classification in such circumstances faces weaker training convergence and poorer signal calling. For the five classes considered here, a passive stabilization approach was used that optimized the kernels for high rejection. More active (computational) stabilization methods are being studied for larger multiclass problems and improved accuracy overall.

Blockade Mechanism For Nine Base-Pair DNA Hairpins

In a forthcoming manuscript (Winters-Hilt et al., in preparation), analysis will focus on details of the current blockade mechanism, so only a preliminary description is given here (Figure 3). The intermediate level (IL) conductance state initiates most blockades and always transitions to the upper level conductance state (UL). This is explained by binding of the hairpin terminus to the vestibule interior (IL) followed by desorption of the DNA from the protein wall and orientation of the stem along the axis of the electric field (UL). Transitions from the UL state were either back to the IL state or to the lower level conductance state (LL). From the LL state there were brief transitions to nearly full blockade, denoted by S for spike conductance state. The LL and S states are both thought to involve binding between the hairpin's terminal 5′ base and the pore's limiting aperture. The brief S state behavior is explained by a terminus-fraying event that is accompanied by extension by the terminal 3′ base into the limiting aperture. Part of the evidence for this is a strong spike (fraying) frequency correlation with the different terminus binding energies. Asymmetric base addition or phosphorylation (at the terminal 3′ and 5′ positions) is part of the evidence for the asymmetric roles for 5′ binding (LL and S) and 3′ fraying/extension (S).

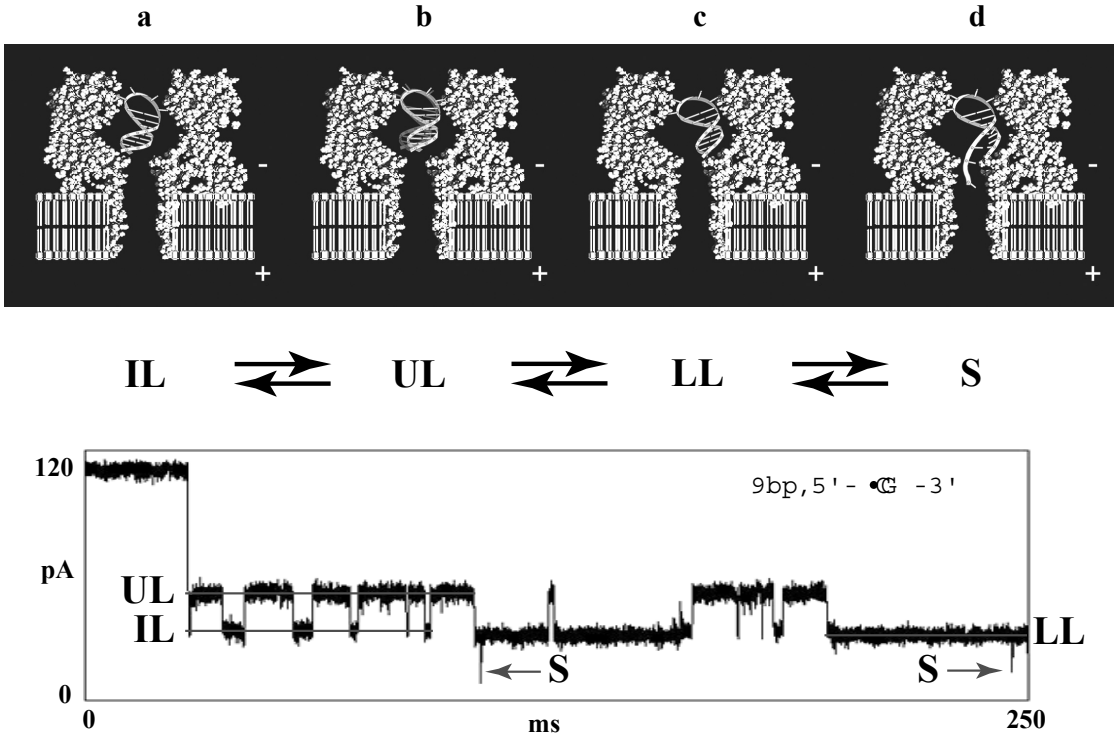

FIGURE 3. Molecular mechanisms underlying the observed current transitions. a) When a 9bp DNA hairpin initially enters the pore, the loop is perched in the vestibule mouth and the stem terminus binds to amino acid residues near the limiting aperture. This results in the IL conductance level. b) When the terminal base pair desorbs from the pore wall, the stem and loop may realign, resulting in a substantial current increase to UL. Interconversion between the IL and UL states may occur numerous times, or UL may convert to the LL state c). The LL state corresponds to binding of the stem terminus to amino acids near the limiting aperture

but in a different manner from IL. d) From the LL bound state, the duplex terminus may fray resulting in extension and capture of one strand in the pore constriction.

Force/Geometry Probing Using DNA Hairpins

In another forthcoming manuscript (DeGuzman et al., in preparation), a variety of DNA hairpins are used as probes of the α-hemolysin protein channel geometry. The same experiments also serve to reveal the forces at various points in the channel. Building on the work of (Vercoutere et al., 2001), a series of blunt-ended DNA hairpins are used to probe the depth of the vestibule (where the dsDNA stems can reach). The blockade signal exhibits a single blockade level for hairpins with stem lengths ranging from three base-pairs (3bps) to seven base-pairs (7bps). For the 8bp hairpin a telegraph signal appears, with the primary blockade level at the greater resistance. For 9bp hairpins, and those with longer stems, there appear to be three main levels (the 9bp case is discussed above). The geometric bottom of the vestibule is reached with a 9bp hairpin, ±1bp. The indeterminacy results from the unknown positions of the binding events thought to correlate with the lower conductance levels and is being studied further. Using the 9bp hairpin as a base, and taking into account the abovementioned 3′-fraying/extension hypothesis, single-stranded DNA overhangs of varying length were added to the base at the 3′ terminus. This permits critical probing through the trans-membrane part of the channel in a very controlled manner, by a single captured molecule event. Preliminary results indicate two significant trans-membrane constrictions, one at the limiting aperture, and one near the trans-opening. The resolving power of the limiting-aperture/trans-opening constrictions is of critical importance in DNA-sequencing and biosensor applications, and is undetermined as of yet.

Sequencing Possibilities

For sequencing, the single molecule basis of nanopore measurements may permit Sanger-type sequencing on DNA molecules separated by capillary electrophoresis. If ssDNA translocation for α-hemolysin can be slowed enough, by use of single-enzyme couplings or servo-electronics, then single-molecule DNA sequencing may prove possible as well. For single-molecule sequencing to be successful, however, the deconvolution problem must be solved for the collection of bases at the main current restrictions (where, presumably, the greatest physical imprint is made on the ionic current). Deconvolution of base content from a single blockade signal may be possible if dominant contributions to resistance span only 20 Å or so (amounting to about three nucleotides length of ssDNA). If the critical nanopore thickness can be made less than 20 Å it may prove advantageous to work with dsDNA. Such an approach would probably require symmetry breaking on the dsDNA strand via base-pairing a native ssDNA strand with a set of Watson-Crick substitute nucleotides. Although dsDNA takes about twice as many bases to cover the same distance, the information imparted in those bases is much richer than that for ssDNA (due to Watson-Crick type bond formation, etc.), and this may prove critical to obtaining a working sequencer,

particularly given the much greater information that can be extracted by introducing excitation to the dsDNA bonds during observation.

Other Applications

One of the key strengths of nanopore detectors is that they analyze populations of single molecules. With signal processing and pattern recognition, this information enables a new type of cheminformatics based on channel current measurements. Single molecule observations are also of interest in biophysics; binding/conformational changes on captured dsDNA end regions, for example, might be tracked and understood using the nanopore blockade signal. DNA regions away from the ends may eventually be studied in a similar manner, using pore-translocation that resolves (and is dominated by) the distinctive conductance/binding imprint of those bases threading the limiting aperture constriction. For single nucleotide polymorphism (SNP) identification, small sample volumes can be used, such that PCR amplification may not be needed. Non-PCR expression analysis, in general, may offer a new method for biological experimentation on live cells using patch-clamp methods.

CONCLUSION

Five species of DNA hairpin were examined, four of which differed only in their terminal base-pairs. Classification of a single 100 ms hairpin event, with no rejection, was 77% accurate on average. Accuracy was boosted above 99% if longer event durations were used or if multiple short events were used with nonzero rejection. For purposes of rapid mixture analysis, the latter approach was adopted, with single species identification with 99.6% accuracy in six seconds, and two species mixture identification in 40 seconds with less than 1% error in the majority species percentage. The signal processing architecture that accomplished this used HMMs for feature extraction and SVMs for classification. The HMMs were implemented with Expectation/Maximization and the SVMs were implemented with novel Kernels. The on-line signal processing was designed to be scalable to hundreds of species, or more, while at the same time performing the classification in less time than the duration of the signal acquisition itself (100 ms). This was accomplished on an inexpensive PC. An unconstrained training process, as used here, has scalability complications due to rapid growth in multiclass combinatorics, but for five species was easily automated (on a network of five PCs). If scalability requirements are relaxed, allowing species-specific HMM processing for example, discrimination accuracy (or speed) can be boosted even further. The processing architecture is directly applicable to other channel current analysis situations by simply re-training the machine learning components.

ACKNOWLEDGEMENTS

I thank my collaborators from the larger exposition of this work (Winters-Hilt et al., 2003): W. Vercoutere, V. S. DeGuzman, D.W. Deamer, M. Akeson, and D. Haussler.

I benefited greatly from discussions of biophysics, biochemistry, and nanopore operation with W. Vercoutere, V. S. DeGuzman, D.W. Deamer, and M. Akeson. Discussions with D. Haussler provided insight and clarity to the development of the machine learning solution. A special thanks to Mark Akeson for a careful reading of early drafts. Thanks also to Sam Ridino and Joseph T. Rodgers for help with data acquisition. This work was funded by the National Human Genome Research Institute and by the Howard Hughes Medical Institute.

REFERENCES

Akeson M, D. Branton, J.J. Kasianowicz, E. Brandin, D.W. Deamer. 1999. Microsecond Time-Scale Discrimination Among Polycytidylic Acid, Polyadenylic Acid, and Polyuridylic Acid as Homopolymers or as Segments Within Single RNA Molecules. Biophys. J. 77(6):3227-3233.

Bayley, H. 2000. Pore planning: Functional membrane proteins by design. J. Gen. Physiol. 116. 1a.

Bredensteiner, E.J.; and K.P. Bennett. 1999. Multicategory classification by support vector machines. Comput. Optim. and Appl. 12. 53-79.

Burges, C.J.C**.** 1998. A tutorial on support vector machines for pattern recognition. Data Min. Knowl. Discov., 2. 121-67.

Chung, S.H., J.B. Moore, L. Xia, L. S. Premkumar, and P. W. Gage. 1990. Characterization of single channel currents using digital signal processing techniques based on Hidden Markov models. Phil. Trans. R. Soc. Lond. B 329. 265-285.

Chung, S-H., and P. W. Gage. 1998. Signal processing techniques for channel current analysis based on hidden Markov models. *In* Methods in Enzymology; Ion channels, Part B. P. M. Conn editior. Academic Press, Inc., San Diego. 420-437.

Colquhoun, D., and F. J. Sigworth. 1995. Fitting and statistical analysis of single-channel products. *In* Single-channel recording. B. Sakmann and E. Neher editors. Second edition. Plenum Publishing Corp., New York. 483-587.

Cormen, T.H., C. E. Leiserson, and R. L. Rivest. 1989. Introduction to Algorithms. MIT-Press, Cambridge, USA.

Cover, T. M. and J. A. Thomas. 1991. Elements of Information Theory. Wiley-Interscience, New York.

Diserbo M, P. Masson, P. Gourmelon, and R. Caterini, 2000. Utility of the wavelet transform to analyze the stationarity of single ionic channel recordings. J. Neurosci. Methods 99(1-2):137-141.

Durbin R. 1998. Biological sequence analysis : probalistic models of proteins and nucleic acids. Cambridge, UK New York: Cambridge University Press. xi, 356 p.

Gouaux J.E., O. Braha, M.R. Hobaugh, L. Song, S. Cheley, C. Shustak, and H. Bayley, 1994. Subunit stoichiometry of staphylococcal alpha-hemolysin in crystals and on membranes: a heptameric transmembrane pore. Proc. Natl. Acad. Sci. USA 91:12828-12831.

Jaakkola, T. S., and D. Haussler. 1998. Exploiting generative models in discriminative classifiers. *In* Advances in Neural Processing Systems 11. Cambridge, MA, 1999. MIT Press.

Joachims, T. 1998. Making large-scale SVM learning practical. *In* Advances in Kernel Methods -- Support Vector Learning. B. Scholkopf, C. J. C. Burges, and A. J. Smola editors. MIT Press, Cambridge, USA. Ch. 11.

Kasianowicz, J.J., E. Brandin, D. Branton, and D.W. Deamer, 1996. Characterization of Individual Polynucleotide Molecules Using a Membrane Channel. Proc. Natl. Acad. Sci. USA 93(24):13770-13773.

Li, J., C. McMullan, D. Stein, D. Branton, and J. Golovckenko. 2001. Solid state nanopores for single molecule detection. Biophys. J. 80. 339a.

Meller A, L. Nivon, E. Brandin, J. Golovchenko, and D. Branton, 2000. Rapid nanopore discrimination between single polynucleotide molecules. Proc. Natl. Acad. Sci. USA 97(3):1079-1084.

Meller A, L. Nivon, and D. Branton, 2001. Voltage-driven DNA translocations through a nanopore. Phys. Rev. Lett. 86(15):3435-3438.

Osuna, E.; R. Freund, and F. Girosi. 1997. An improved training algorithm for support vector machines. *In* Neural Networks for Signal Processing VII. J. Principe, L. Gile, N. Morgan, and E. Wilson editors. IEEE, New York. 276-85.

Platt, J. C. 1998. Fast Training of Support Vector Machines using Sequential Minimal Optimization. *In* Advances in Kernel Methods -- Support Vector Learning. B. Scholkopf, C. J. C. Burges, and A. J. Smola editors. MIT Press, Cambridge, USA. Ch. 12.

SantaLucia J. 1998. A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. Proc. Natl. Acad. Sci. USA 95(4):1460-1465.

Song L., M.R. Hobaugh, C. Shustak, S. Cheley, H. Bayley, and J.E. Gouaux, 1996. Structure of Staphylococcal Alpha-Hemolysin, a Heptameric Transmembrane Pore. Science 274 (5294):1859-1866.

Vapnik, V. N. 1999. The Nature of Statistical Learning Theory (2nd ed.). Springer-Verlag, New York.

Vercoutere W., S. Winters-Hilt, H. Olsen, D.W. Deamer, D. Haussler, and M. Akeson, 2001. Rapid discrimination among individual DNA hairpin molecules at single-nucleotide resolution using an ion channel. Nat. Biotechnol. 19(3):248-252

Winters-Hilt, S., W. Vercoutere, V. S. DeGuzman, D.W. Deamer, M. Akeson, and D. Haussler, 2003. Highly Accurate Classification of Watson-Crick Base-Pairs on Termini of Single DNA Molecules. Biophys. J. 84 (2) 1-10.