DNA AND CELL BIOLOGY
Volume 23, Number 10, 2004

Pp. 000–000

Nanopore Cheminformatics

STEPHEN WINTERS-HILT,[1,2] and MARK AKESON[3,4]

ABSTRACT

A cheminformatics method is described for classification, and biophysical examination, of individual molecules. A novel molecular detector is used—one based on current blockade measurements through a nanometer-scale ion channel (alpha-hemolysin). Classification results are described for blockades caused by DNA molecules in the alpha-hemolysin nanopore detector, with signal analysis and pattern recognition performed using a combination of methods from bioinformatics and machine learning. Due to the size of the alpha-hemolysin protein channel, the blockade events report on one DNA molecule at a time, which enables a variety of reproducible, single-molecule biophysical experiments. To capture the full sensitivity of the nanopore detector's blockade signal, Hidden Markov Models (HMMs) were used with Expectation/Maximization for denoising and for associating a feature vector with the ionic current blockade of each captured DNA molecule. Support Vector Machines (SVMs) that employ novel kernel designs were then used as discriminators. With SVM training performed off-line, and economical HMM processing on-line, blockade classification was possible during capture. HMMs were also used in conjunction with a time-domain finite state automaton (off-line) for feature discovery and kinetics analysis. Analysis of the DNA data indicates a variety of binding (DNA–protein), fraying, and conformational shifts that are consistent with data obtained from thermodynamic analyses (melting curves), X-ray crystallography, and NMR studies. The software tools are designed for analysis of generic blockades in ionic channels, including those in other biological pore-forming toxins, other biological channels in general, and semiconductor-based channels.

INTRODUCTION

A NANOMETER-SCALE CHANNEL can be used to associate ionic current measurements with single-molecule channel blockades. A biologically based (protein) channel, α-hemolysin, is used for this purpose because its solution soluble monomer easily self-assembles in membranes as a heptamer channel (Gouaux *et al*., 1994; Song *et al*., 1996). This leads to an inexpensive and reproducible nanopore detector (see Fig. 1a). α-Hemolysin is also chosen because it is stable (e.g., nongating) and has dimensions well suited to DNA/RNA measurement: ssDNA translocates while dsDNA does not, being held in the channel's *cis*-side vestibule instead. Figure 1b shows a crystal structure (Song *et al*., 1996), with a 1.5-nm limiting aperture ringed by Glutamic Acids and Lysines. The entry aperture on the *cis*-side is 2.6 nm in diameter (ringed by Threonines), which is large enough to admit (and capture) dsDNA. Figure 1b shows a captured nine base-pair DNA hairpin superimposed. Operation of the α-hemolysin nanopore detector demonstrates that it is possible to obtain at least Angstrom-level resolution of structural features (Winters-Hilt *et al*., 2003). To accomplish this, however, the detector must extract subtle differences between ionic current blockades Figure 2 shows blockade traces for a collection of five hairpins. Figure 2b shows the dominant blockades, and their frequencies, for the different hairpin molecules.

A nanometer-scale, α-hemolysin based, channel detector, or "nanopore detector," can be used to observe single ssDNA molecules during channel translocation (Kasianowicz *et al*., 1996; Akeson *et al*., 1999; Meller *et al*., 2000, 2001), or to observe the ends of single dsDNA molecules captured by the pore (Vercoutere *et al*., 2001, Winters-Hilt et al., 2003). For the α-hemolysin nanopore detector, progress analyzing ssDNA translocations has been limited due to the high speed of such translations. Lowering the applied potential to slow the ssDNA

[1]Department of Computer Science, University of New Orleans, New Orleans, Louisiana.
[2]Research Institute for Children, Children's Hospital, New Orleans, Louisiana.
[3]Department of Chemistry and Biochemistry, University of California, Santa Cruz, California.
[4]Howard Hughes Medical Institute, University of California, Santa Cruz, California.

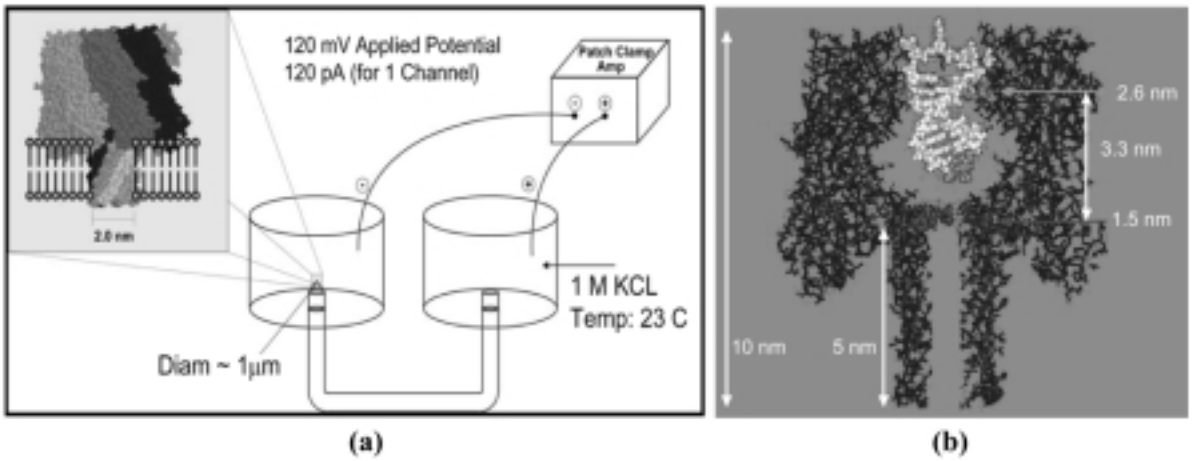

FIG. 1. The nanopore detector. (**a**) The electrochemistry setup for the nanopore device. (**b**) The crystallographic description of the α-hemolysin channel with a 9-bp DNA hairpin superimposed.

translocation does not help either, since a minimal applied potential is required to draw molecules into the channel, that is, a free energy barrier must be overcome. For end-capture of ds-DNA, on the other hand, extensive characterization of ionic current blockades is possible because the molecules can be held and observed for as long a needed. A voltage-reversal sampling cycle then allows examination of many such dsDNA ends (Vercoutere *et al*., 2001; Winters-Hilt *et al*., 2003). Modifications to the α-hemolysin channel have been examined (Bayley, 2000), and semiconductor nanopores are being developed (Li *et al*., 2001). In previous work (Winters-Hilt *et al*., 2003) it was shown that molecular blockade information permitted highly accurate classification of DNA hairpins (99.6% accuracy, see Figs. 3 and 4). The information that permitted this discrimination was found to derive from an imprint of the DNA–protein binding kinetics on the surrounding ionic flow. In this paper, preliminary results show that a nanopore detector, coupled with modern pattern recognition methods, can also be used to characterize the conformational kinetics of captured DNA hairpins.

In the nanopore signal analysis in (Winters-Hilt *et al*.,) (see Fig. 3), a Hidden Markov Model (HMM) was used to extract a feature vector from each blockade example. HMMs can characterize current blockades by identifying a sequence of sub-blockades as a sequence of state emissions (Chung *et al*., 1990; Colquhoun and Sigworth, 1995; Chung and Gage, 1998). The parameters of an HMM can then be estimated using a method

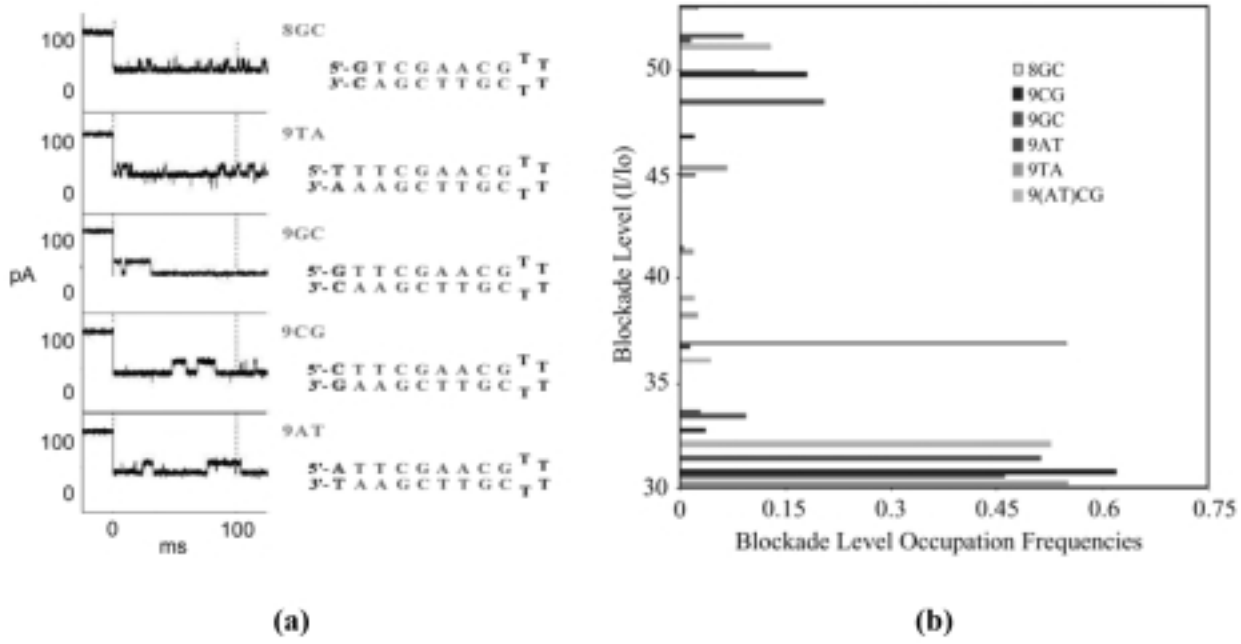

FIG. 2. The channel current blockade signal. (**a**) The five DNA hairpins, with sample blockades, that were used to test the sensitivity of the nanopore device. (**b**) The dominant blockades, and their frequencies, for the different hairpin molecules.

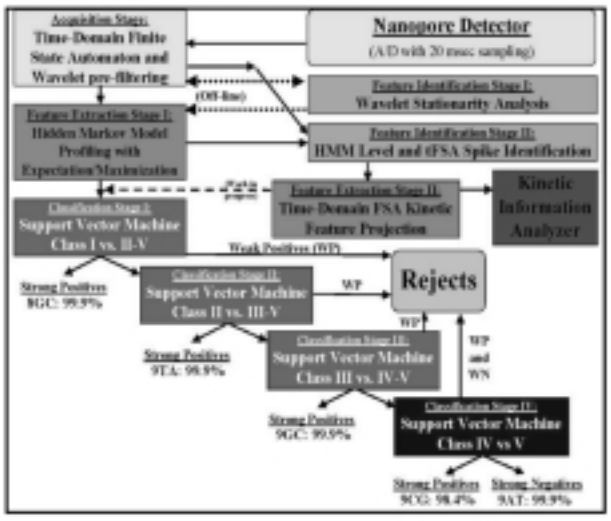

FIG. 3. The signal processing architecture. Signal acquisition was performed using a time-domain, thresholding, Finite State Automaton. This was followed by adaptive prefiltering using a wavelet-domain Finite State Automaton. Feature extraction on those acquired channel blockades was done by Hidden Markov 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 multiclass SVM architecture benefits from strong signal skimming and weak signal rejection along the line of decision nodes. Scalability to larger multiclass 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 15th single-molecule sampling/classification (in approx. 6 sec). Off-line kinetic feature extraction was done at the newly added HMM Level and tFSA Spike Identification module and the time-domain FSA Kinetic Feature Projection module.

called Expectation/Maximization (Durbin, 1998). Although HMMs can be used to discriminate among several classes of input, multiclass computational scalability tends to favor their use as feature extractors. This, and related signal processing issues, can be found in Winters-Hilt *et al*. (2003). Thus, HMMs, which are well suited to extraction of aperiodic information embedded in stochastic sequential data, are used for feature extraction. Classification of feature vectors obtained by the HMM (for each individual blockade event) is then done using Support Vector Machines (SVMs), an approach which automatically provides a confidence measure on each classification. SVMs are fast, easily trained, discriminators (Burges, 1998; Vapnik, 1998), for which strong discrimination is possible without the overfitting complications common to neural net discriminators (Vapnik, 1998).

AU1

In this work, the signal processing architecture shown in Figure 3 has grown to include an added feature identification module (identification stage II) and an added feature extraction module (extraction stage II). The purpose of these added computational stages is to build on the feature identification/extraction information available from the stage I modules, such that kinetic information can be directly extracted and represented. This information will eventually be used for on-line signal processing (shown as the dotted arrow in Fig. 3). Such on-line processing will require a new feature vector structure and retraining of the SVM Decision Tree, however, so the focus in this work is on obtaining and examining preliminary results off-line (via the "Kinetic Analyzer" module in Fig. 3).

Prior classification and mechanism results

Five DNA hairpins were used in the prototype study to explore the sensitivity of the detector, as well as to probe the pore geometry (see Discussion). The DNA hairpins studied (in Fig. 2a) differed only in their terminal base pairs. Classification accuracy was 99.6% on average for the five DNA hairpins (see Fig. 4a), and this was accomplished by the 15th classification attempt (6 sec on average). The classification result for a mixture solution of 9TA and 9GC hairpin species (Fig. 4b) is shown as the number of single molecule samplings is increased. The mixture was in a 3:1 ratio of 9TA:9GC, consistent with the 75% asymptote. Less than 1% error (on majority population size) was obtained by the 100th observation.

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 (see Fig. 2b). 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 UL blockade level. Figure 5 describes the hypothesized blockade mechanism for the nine base-pair hairpin blockades (Vercoutere *et al*., 2003; Winters-Hilt *et al*., 2003 for further results).

F5

Preliminary results indicate that the UL blockade level may be unbound at its terminus, permitting conformational kinetics to be seen. One example of this is that the upper level blockade (UL) plateaus once the hairpin stem length reaches seven base pairs. This plateau occurs well before that of the other blockade levels (which can be explained as the hairpin growing too long for the pore vestibule, causing the hairpin loop to extend beyond the pore vestibule entrance). The explanation for the UL plateau centers on the tight flow geometry between channel and captured hairpin. In such a geometry, much of the ionic flow is confined to be in or near the grooves of the captured DNA molecule. For the unbound molecule, this groove flow can be directed towards the limiting aperture by appropriate orientation of the hairpin molecule (which it is free to do since it is unbound). The unbound molecule can thus cause a "short circuit" effect, where the contribution to the ionic current is not significantly altered as the hairpin is extended (by base-pair addition), thus explaining the early plateau. This, and other, results (described in Winters-Hilt *et al*., 2003 and Vercoutere et al., 2003) strengthen the hypothesis that the nine base-pair DNA hairpin's UL blockade corresponds to a molecular state with unbound terminus.

MATERIALS AND 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-mi-

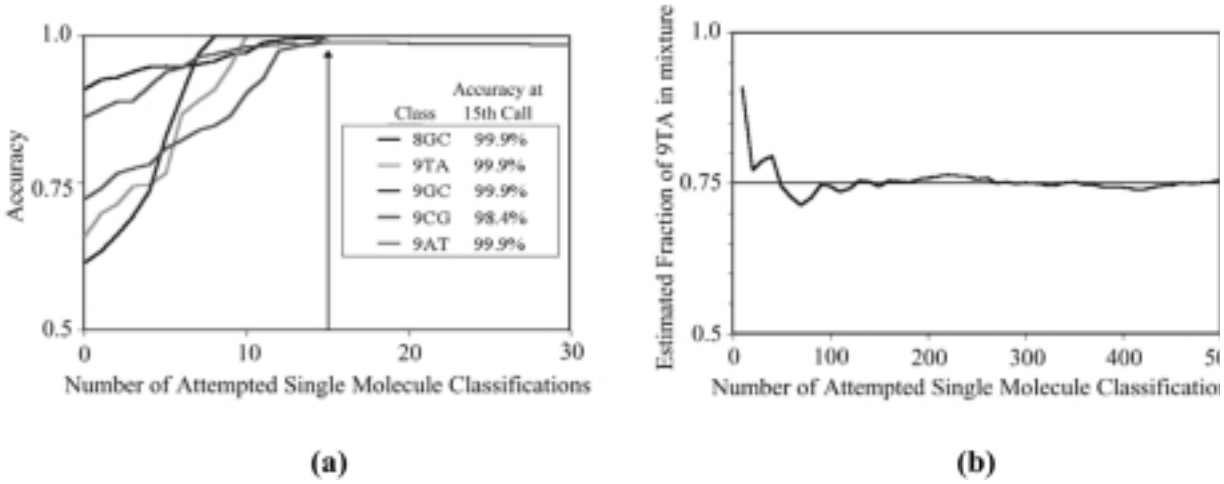

FIG. 4. Single-species and mixture classification results. (**a**) The prediction accuracy as the number of signal classification attempts increases (allowing increase in the rejection threshold). (**b**) The prediction accuracy on 3:1 mixture of 9TA to 9GC DNA hairpins.

cron diameter horizontal Teflon aperture (Vercoutere *et al*., 2001). The bilayer separates two 70-μl 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, that is, 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′ CTTCGAACG-TTTTCGTTCGAAG 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 *c*is 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-μsec 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 generic signal identification/acquisition for the first 100 msec of blockade signal (Acquisition Stage, Fig. 3). An abrupt drop to 70% residual current, or less, triggered transition from the reset ready state to the signal active state. For DNA hairpins with stems shorter than eight base pairs, multiple states were not clearly discernible by the prototype, presumably because the hairpins were too short to interact with the current constriction and strong forces near the limiting aperture. For nine base-pair hairpins, and longer, a clear 1/f noise (flicker noise) is discernible—a preliminary indication of the single-molecule kinetics results that follow. The effective duty cycle for acquiring the desired 100-msec blockade measurements was one reading every 0.4 sec. Further details on the voltage toggling protocol and the time-domain FSA are in Winters-Hilt *et al*., (2003).

Signal preprocessing and unsupervised feature extraction

Each 100-msec signal acquired by the time-domain FSA consisted of a sequence of 5000 subblockade levels (with the 20-μsec 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, Fig. 3; also see 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. 3). The HMM was implemented with 50 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 denoised by five rounds of Expectation-Maximization (EM) training on the parameters of the HMM. After the EM iterations, 150 parameters were extracted from the HMM. The 150 feature vector components were extracted from parameterized emission probabilities, a compressed representation of transition probabilities,

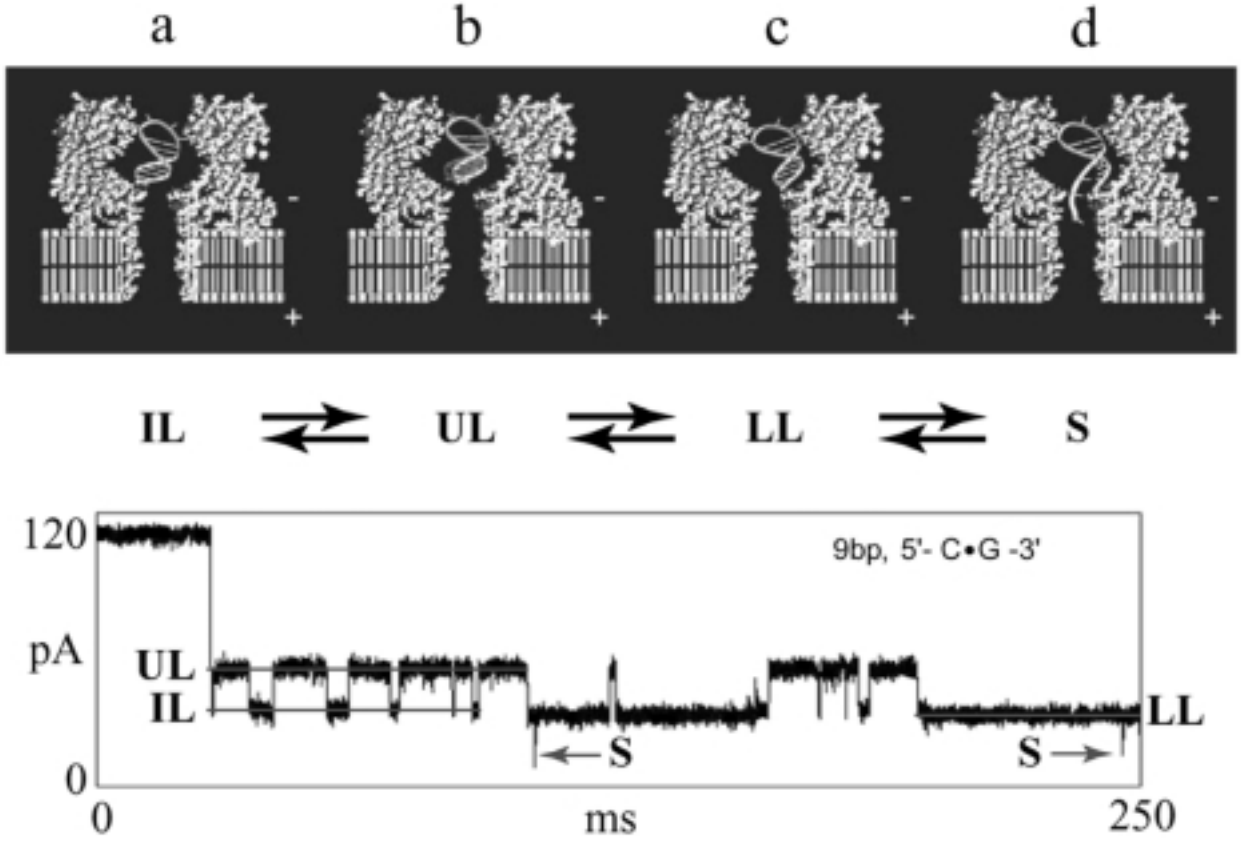

FIG. 5. The nine base pair DNA hairpin blockade mechanism. Molecular mechanisms underlying the observed current transitions. (**a**) When a 9-bp DNA hairpin initially enters the pore, the loop is perched in the vestibule mouth and the stem terminus binds to amino acid residues near the limiting aperture. 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.

and use of *a posteriori* information deriving from the Viterbi path solution (further details in Winters-Hilt *et al*., 2003). This information elucidates the blockade levels (states) characteristic of a given molecule, and the occupation probabilities for those levels (Fig. 2b), but does not directly provide kinetic information. 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.

Kinetic feature extraction

Extraction of kinetic information begins with identification of the main blockade levels for the various blockade classes (off-line). This information is then used to scan through already labeled (classified) blockade data, with projection of the blockade levels onto the levels identified for that class of molecule. A time-domain FSA performs the above scan, and uses the information obtained to tabulate the lifetimes of the various blockade levels. Once the lifetimes of the various levels are obtained, a variety of kinetic properties can be obtained. If the experiment is repeated over a range of temperatures, a full set of kinetic data is obtained. This data can be used to calculate k_{on} and k_{off} rates for binding events, as well as indirectly calculate forces by means of the van't Hoff Arrhenius equation (or other arguments based on Boltzmann factors).

Classification training

The normalized feature vectors obtained from the feature extraction stage are 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 the greatest training-instance void, or "margin," around the decision hyperplane. Binary SVMs can be grouped into a classifier tree to perform multiclass discrimination, and this was done here for the five classes of DNA hairpin (shown in classification stages I–IV in Fig. 3). Tuning on the multiclass SVM architecture itself was done for performance optimization, and separate tuning was 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 and Haussler, 1998) and regularized information divergences. In the former case, the squared

Euclidean distance between feature vectors **x** and **y,** $d^2(\mathbf{x},\mathbf{y}) = \Sigma_k(\mathbf{x_k} - \mathbf{y_k})^2$, also known as the squared l_2-norm on (**x-y**), $[l_2(\mathbf{x}\text{-}\mathbf{y})]^2 = d^2(\mathbf{x},\mathbf{y})$, is associated with the Gaussian kernel: $K_G(\mathbf{x},\mathbf{y}) = \exp(-d^2(\mathbf{x},\mathbf{y})/2\sigma^2)$. The latter case represents a whole new class of kernels (see Winters-Hilt *et al*., 2003, for more details) based on information-theoretic measures of distance between probability vectors (discrete distributions). The information divergence (relative entropy) between probability vectors **x** and **y,** $\boldsymbol{D}(\mathbf{x}||\mathbf{y}) = \Sigma_k \mathbf{x}_k \log(\mathbf{x}_k/\mathbf{y}_k)$, can be associated with the "Entropic kernel:" $\boldsymbol{K_E}(\mathbf{x},\mathbf{y}) = \exp(-[D(\mathbf{x}||\mathbf{y}) + D(\mathbf{y}||\mathbf{x})]2\sigma^2)$. The terminating SVM node of the classifier tree (stage IV in Fig. 1) performed best with such an Entropic kernel. The other nodes of the classifier tree used a regularized-distance type kernel, the "Variation-distance kernel," based on the square root of the l_1-norm, where $l_1(\mathbf{x}\text{-}\mathbf{y}) = \Sigma_k|\mathbf{x_k} - \mathbf{y_k}|$, with kernel $K_l(\mathbf{x},\mathbf{y}) = \exp(-\sqrt{l_1(\mathbf{x}\text{-}\mathbf{y})}/2\sigma^2)$.

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 multiclass SVM training was based on over 10,000 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).

Prototype testing protocol

In the five DNA hairpin study, the test data consisted of over 2000 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

The prototype study described in the introduction indicates that the UL blockade state (see description with Fig. 5) can be understood as a captured DNA hairpin with unbound terminal base pair. This enables a study of the conformational dynamics at the ends of DNA molecules by focusing on the UL states of nine base-pair DNA hairpins. In preliminary results, shown in Figure 6, it can be seen that the UL blockade state for some of the hairpin molecules actually has internal structure.

F6

This is currently hypothesized to be due to conformational switching in the hairpin stem. DNA hairpins that do not exhibit such switching are thought to exist in one, dominant, helical conformation. Since much of the current flow is thought to reside in the DNA's major groove (see Discussion section), it is understandable why changes in helical conformation might imprint as toggles in the UL blockade level. Information from NMR studies on the same tetranucleotide termini confirms the results indicated, for example, the molecule thought to be exhibiting conformational switching in Figure 6 is found in NMR studies to have two low energy states.

DISCUSSION

Nanopore cheminformatics provides a powerful new tool for single molecule biophysics. Preliminary efforts indicate that a variety of sequencing and other biotechnology schemes will be possible. Likewise, nanopore-based cheminformatics offers an exciting new arena in which to develop and test the latest machine learning approaches. So far, every machine-learning method introduced, including HMMs and SVMs, has enabled greater sensitivity to be extracted from the nanopore device.

Major groove ion flow

Given the restricted flow geometry between protein channel and a captured DNA hairpin, it is perhaps surprising that a number of unexpected nanomechanical and nanofluidic issues have not arisen. So far, there is only the odd "short circuit" effect described earlier. Further study of conformational switching will inevitably have to address some of these issues, since they are observable in precisely the odd state referred to above. One interesting possibility along these lines is that of cooperative flow along the major groove the DNA molecule.

λ-Exonuclease as a brake on ssDNA translocation

Using lambda exonuclease as an ssDNA brake appears to be possible. Conditions have been obtained where both the exonuclease retains function and the toxin self-assembles. Other work on methylated or dye-tagged ssDNA and dsDNA appears to offer significant new information as well (without laser excitation of dyes being introduced yet). Experiments with laser modulation of analytes are in progress.

Force/geometry probing using DNA hairpins

For a forthcoming manuscript, 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. This is done by 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. The blockade signal exhibits a single blockade level for hairpins with stem lengths ranging from three base pairs (3 bp) to seven base pairs (7 bp). For the 8-bp hairpin a telegraph signal appears, with the primary blockade level at the greater resistance. For 9-bp hairpins, and those with longer stems, there appear to be three main levels. The geometric bottom of the vestibule is reached with a 9-bp hairpin, ± 1 bp. Using the 9 bp hairpin as a base, and

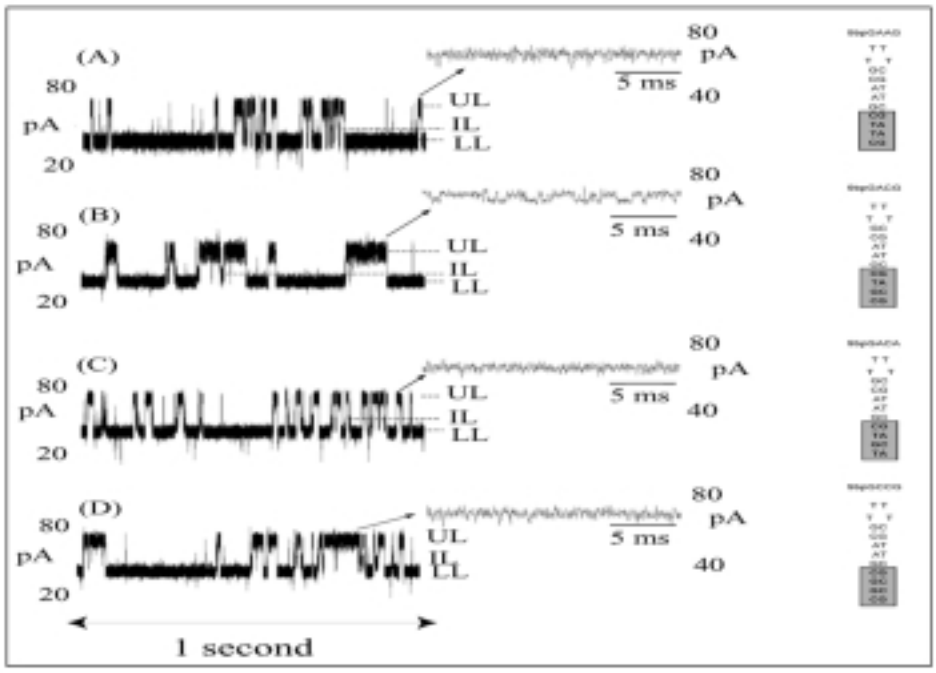

FIG. 6. UL toggle correlates with NMR predicted conformational switching. Preliminary results add credence to the hypothesis that the UL state has an unbound terminus, and that while in that state conformational switching may be observable. For the molecules and blockades shown, three are found by NMR to have one dominant ground state (no switching), while one of the molecules is found by NMR to have two dominant low energy states (switching). This corresponds exactly with what has been observed in terms of the existence of fine structure (toggling) in the UL blockade state.

taking into account the 3′-fraying/extension hypothesis (mentioned in the mechanism description in Fig. 5), single-stranded DNA overhangs of varying length were added to the base at the 3′ terminus. This permits critical force/geometry probing of 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

The highly accurate "read" on the end of nine base pair DNA hairpins appears to extend to 10, 11, and 12 base pair hairpins as well. Thus, the possibility of performing a similar "read" on the end of native blunt-ended dsDNA seems possible as well. In conjunction with capillary electrophoresis, this offers the prospect of the entire Sanger sequencing protocol being performed on a microchip-sized laboratory. If ssDNA translocation through 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). Thus, single-molecule sequencing will require further progress in the force/geometry probing and the enzyme braking efforts. Since dsDNA carries much more information than ssDNA (i.e., the molecular motions are much more constrained and readable), progress may eventually be made with easily formed synthetic/ssDNA chimeric molecules that are sized more like ssDNA, but have the richer bond-formation structure of dsDNA.

Non-PCR expression analysis

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. For single nucleotide polymorphism (SNP) identification, a nanopore detector also offer the prospect that only small sample volumes need be used, such that PCR amplification may not even be needed. Non-PCR expression analysis, in general, may offer a new method for biological experimentation on live cells using patch-clamp methods.

Novel kernels

The kernels studied were not limited to those satisfying Mercer's conditions. The variation and entropic kernels, however, probably satisfy Mercer's conditions, since they can be described as metrics "regularized" by incorporation as positive arguments in a decaying exponential. The Gaussian kernel, which satisfies Mercer's conditions, has the exponential form (with Euclidean distance squared between feature vectors) and was

outperformed in all cases studied by the entropic and variation kernels. The original motivation for working with the entropic kernel was to obtain a faster, more noise-resistant kernel for information obtained via an HMM feature extractor (instead of the theoretically attractive choice of directly integrating the two via a Fisher Kernel (Jaakkola and Haussler, 1998)). This led to a general formulation where feature extraction was designed to arrive at probability vectors (i.e., discrete probability distributions) on a predefined, and complete, space of possibilities. (The different blockade levels, and their frequencies, for example.) This turns out to be a very general formulation, wherein feature extraction makes use of signal decomposition into a complete set of separable states. A probability vector formulation also provides a straightforward hand-off to the SVM classifiers, since all feature vectors have the same length with such an approach. What this means for the SVM is that geometric notions of distance are no longer the best measure for comparing feature vectors. For probability vectors (i.e., discrete distributions), the best measures of similarity are the various information-theoretic divergences: Kullback-Leibler, Renyi, etc. By symmetrizing over the arguments of those divergences we obtain a rich source of kernels that might work well with the types of probabilistic data obtained. Thus far, only the Kullback-Leibler divergence has been extensively studied in this manner (giving rise to the entropic kernel).

A multiclass discriminator can be implemented using binary SVMs grouped in a decision tree architecture (as in Fig. 3). Alternatively, a (single) multiclass SVM can be implemented. The latter takes on a much more complicated form that appears much more susceptible to noise, however, and is much more difficult to train since larger "chunks" are needed to carry all the support vectors. Although the monolithic SVM approach is clearly not scalable, it may offer better performance when working with small class sets. The monolithic approach also avoids the combinatorial explosion caused when optimizing a decision tree architecture. It was revealed in Winters-Hilt *et al*., (2003), however, that the SVM's rejection capability often leads to the optimal architecture reducing to a linear tree architecture with strong signals skimmed off class by class.

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 at the worst place in an on-line pattern recognition setting: the expensive feature extraction stage. In ongoing work, semiscalable, 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 (the kinetic data features). For the kinetic-type data analysis, this reduced feature set may be obtainable, reliably, via an unsupervised, scalable, learning process (shown as the dotted arrow in Fig. 3). The extent to which this is possible is being explored this time.

Passive versus active signal stabilization

Reestablishing 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 in the prototype, a passive stabilization approach was used that optimized the kernels for high rejection. More active (computationally based) stabilization methods are being studied for larger multiclass problems and improved accuracy overall. These methods entail incorporation of control molecules into the experiment (like the eight base-pair hairpin in Fig. 2a) that are tracked as they are randomly sampled along with the analytes of interest.

CONCLUSION

Nanopore cheminformatics based on the α-hemolysin nanopore detector offers a new method for single molecule experimentation. Molecules can be classified by characterization of their binding kinetics and dissociation kinetics (i.e., terminal base pair "breathing" rates). The new, preliminary, result indicated here is that conformational kinetics may be observable as well. On the signal analysis side of this experiment, there is one critical linkage remaining: a link between the kinetic feature information and a retrained SVM decision tree.

ACKNOWLEDGMENTS

For their many helpful conversations and technical assistance, we thank our colleagues at the University of California, Santa Cruz: Veronica DeGuzman, David Deamer, Andrea Solbrig, and Clarence Lee. We also thank our colleagues at the University of New Orleans and the Research Institute for Children, New Orleans: Andrew Duda and Seth Pincus. This work was funded by the Research Institute for Children, New Orleans, the University of New Orleans, the National Human Genome Research Institute, and by the Howard Hughes Medical Institute.

REFERENCES

AKESON, M., BRANTON, D., KASIANOWICZ, J.J., BRANDIN, E., and DEAMER, D.W. (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,** 3227–3233.

BAYLEY, H. (2000). Pore planning: Functional membrane proteins by design. J. Gen. Physiol. **116,** 1a.

BURGES, C.J.C. (1998). A tutorial on support vector machines for pattern recognition. Data Min. Knowl. Discov. **2,** 121–167.

CHUNG, S.-H., and GAGE, P.W. (1998). Signal processing techniques for channel current analysis based on hidden Markov models. In *Methods in Enzymology; Ion Channels*, Part B. P.M. Conn, ed. (Academic Press, Inc., San Diego) pp. 420–437.

CHUNG, S.H., MOORE, J.B., XIA, L., PREMKUMAR, L.S., and GAGE, P.W. (1990). Characterization of single channel currents using digital signal processing techniques based on Hidden Markov models. Philos. Trans. R. Soc. Lond. B **329,** 265–285.

COLQUHOUN, D., and SIGWORTH, F.J. (1995). Fitting and statistical analysis of single-channel products. In *Single-Channel Recording*. B. Sakmann and E. Neher, eds. 2nd ed. (Plenum Publishing Corp., New York) pp. 483–587.

CORMEN, T.H., LEISERSON, C.E., and RIVEST, R.L. (1989). *Introduction to Algorithms*. (MIT Press, Cambridge).

DISERBO, M., MASSON, P., GOURMELON, P., and CATERINI, R. (2000). Utility of the wavelet transform to analyze the stationarity of single ionic channel recordings. J. Neurosci. Methods **99,** 137–141.

DURBIN, R. (1998). *Biological Sequence Analysis: Probalistic Models of Proteins and Nucleic Acids*. (Cambridge University Press, New York) p. xi.

GOUAUX, J.E., BRAHA, O., HOBAUGH, M.R., SONG, L., CHELEY, S., SHUSTAK, C., and BAYLEY, H. (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 HAUSSLER, D. (1998). Exploiting generative models in discriminative classifiers. In *Advances in Neural Processing Systems 11* (MIT Press, Cambridge, MA).

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, eds. (MIT Press, Cambridge, MA), Chap 11.

KASIANOWICZ, J.J., BRANDIN, E., BRANTON, D., and DEAMER, D.W. (1996). Characterization of individual polynucleotide molecules using a membrane channel. Proc. Natl. Acad. Sci. USA **93,** 13770–13773.

LI, J., MCMULLAN, C., STEIN, D., BRANTON, D., and GOLOVCKENKO, J. (2001). Solid state nanopores for single molecule detection. Biophys. J. **80,** 339a.

MELLER, A., NIVON, L., and BRANTON, D. (2001). Voltage-driven DNA translocations through a nanopore. Phys. Rev. Lett. **86,** 3435–3438.

MELLER, A., NIVON, L., BRANDIN, E., GOLOVCHENKO, J., and BRANTON, D. (2000). Rapid nanopore discrimination between single polynucleotide molecules. Proc. Natl. Acad. Sci. USA **97,** 1079–1084.

OSUNA, E., FREUND, R., and GIROSI, F. (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, eds. (IEEE, New York) pp. 276–285.

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, eds. (MIT Press, Cambridge, MA), Chap. 12.

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

SONG, L., HOBAUGH, M.R., SHUSTAK, C., CHELEY, S., BAYLEY, H., and GOUAUX, J.E. (1996). Structure of staphylococcal alpha-hemolysin, a heptameric transmembrane pore. Science **274,** 1859–1866.

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

VERCOUTERE, W., WINTERS-HILT, S., DEGUZMAN, V.S., DEAMER, D.W., RIDINO, S., RODGERS, J.T., OLSEN, H., MARZIALI, A., and AKESON, M. (2003). Discrimination among individual watson-crick base-pairs at the termini of single DNA hairpin molecules. Nucleic Acids. Res. **31,** 1311–1318.

VERCOUTERE W., WINTERS-HILT, S., OLSEN, H., DEAMER, D.W., HAUSSLER, D., and AKESON, M. (2001). Rapid discrimination among individual DNA hairpin molecules at single-nucleotide resolution using an ion channel. Nat. Biotechnol. **19,** 248–252.

WINTERS-HILT, S., VERCOUTERE, W., DEGUZMAN, V.S., DEAMER, D.W., AKESON, M., and HAUSSLER, D. (2003). Highly Accurate classification of Watson-Crick base-pairs on termini of single DNA molecules. Biophys. J. **84,** 1–10.

Address reprint requests to:
Stephen Winters-Hilt
200 Henry Clay Ave.
Research & Education
New Orleans, LA 70118

E-mail: winters@cs.uno.edu

AU2

WINTERS-HILT

AU1
Vapnik 1999?

AU2
degree for corr au? PhD? MD?