Highly Accurate Classification of Watson-Crick Basepairs on Termini of Single DNA Molecules

Stephen Winters-Hilt,*†§ Wenonah Vercoutere,‡ Veronica S. DeGuzman,*‡ David Deamer,‡ Mark Akeson,*§ and David Haussler*†§

*Center for Biomolecular Science and Engineering, University of California, Santa Cruz, California 95064; †Computer Science Department, University of California, Santa Cruz, California 95064; ‡Department of Chemistry and Biochemistry, University of California, Santa Cruz, California 95064; and §Howard Hughes Medical Institute, University of California, Santa Cruz, California 95064

ABSTRACT We introduce a computational method for classification of individual DNA molecules measured by an α-hemolysin channel detector. We show classification with better than 99% accuracy for DNA hairpin molecules that differ only in their terminal Watson-Crick basepairs. Signal classification was done in silico to establish performance metrics (i.e., where train and test data were of known type, via single-species data files). It was then performed in solution to assay real mixtures of DNA hairpins. Hidden Markov Models (HMMs) were used with Expectation/Maximization for denoising 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 multiclass 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.

INTRODUCTION

Molecular classification using nanopore detectors holds promise in biophysics and biotechnology (Akeson et al., 1999; Kasianowicz et al., 1996; Meller et al., 2000; Meller et al., 2001; Vercoutere et al., 2001). Such detectors use a nanometer-scale pore to relate ionic current blockade measurements to single molecule translocation (Akeson et al., 1999; Kasianowicz et al., 1996; Meller et al., 2000) or to 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 optimal for DNA measurement in that single-stranded DNA (ssDNA) translocates whereas double-stranded DNA (dsDNA) does not, being held instead in a vestibule of the pore (Vercoutere et al., 2001). Modifications to the α-hemolysin channel have been examined (Bayley 2000), and semiconductor nanopores are being developed (Li et al., 2001).

For DNA measurements using nanopores, an important milestone is the ability to rapidly identify individual bases or basepairs in single DNA molecules. One end of double-stranded DNA (dsDNA) can be captured by the α-hemolysin pore and held for an extended period of time (Vercoutere et al., 2001). Extensive characterization of the ionic current blockade associated with such an event is thus made possible. In this report, we show that a nanopore detector coupled with machine learning methods can discriminate with high accuracy between DNA hairpins that differ in only one basepair.

In our nanopore signal analysis, an HMM is used to extract a feature vector from each blockade example. Hidden Markov Models (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. HMMs have also been used to estimate state transition and emission probabilities on sequential data in more general contexts, including genomic analysis (Krogh et al., 1994; Stormo, 2000) and voice recognition (Jelinek, 1997). 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, multiclass 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 (for a single blockade event) obtained by the HMM. SVMs are fast, easily trained discriminators (Vapnik, 1999; Burges, 1998). Given a training set of feature vectors, some labeled positive, some labeled negative, SVM training produces an optimized hyperplane that separates the clusters of positives and negatives. Implicit in this is a mapping of feature vectors to points in a higher dimensional space, together with a notion of distance between those points. The distance properties are determined by the choice of kernel in the SVM. Such generality permits strong discrimination, whereas the structural risk minimization that underlies the

Submitted May 30, 2002, and accepted for publication October 15, 2002.

Address reprint requests to Stephen Winters-Hilt, 209 Koshland Way, Santa Cruz, CA 95064. Tel.: 831-420-1395; E-mail: winters@cse.ucsc.edu or haussler@cse.ucsc.edu.

0006-3495/03/02/967/10 $2.00

SVM formulation helps to prevent over-fitting (Vapnik, 1999).

METHODS

Nanopore implementation

Each experiment was conducted using one α-hemolysin channel inserted into a diphytanoyl-phosphatidylcholine/hexadecane bilayer (Fig. 1), where the bilayer was formed across a 20-micron diameter horizontal Teflon aperture (Vercoutere et al., 2001). The bilayer separates two 70-mL chambers containing 1.0 M KCl buffered at pH 8.0 (10 mM HEPES/KOH). A completed bilayer between the chambers was indicated by the lack of ionic current flow when a voltage was applied across the bilayer (using Ag-AgCl electrodes). Once the bilayer was in place, a dilute solution of α-hemolysin (monomer) was added to the *cis* chamber. Self-assembly of the α-hemolysin heptamer and insertion into the bilayer results in a stable, highly reproducible, nanometer-scale channel with a steady current of 120 pA under an applied potential of 120 mV at 23°C (±0.1°C using a Peltier device). Once one channel formed, further pores were prevented from forming by thoroughly perfusing the *cis* chamber with buffer. Molecular blockade signals were then observed by mixing analytes into the *cis* chamber.

DNA hairpin design

The nine basepair hairpin molecules examined in this study share an eight basepair hairpin core sequence, to which we attached one of the four permutations of Watson-Crick basepairs 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′. These are denoted 9GC, 9CG, 9TA, and 9AT. The sequence of the 9CG hairpin was 5′ CTTCGAACGTTTTCGTTCGAAG 3′. The basepairing region is underlined. An eight basepair DNA hairpin with a 5′-G•C-3′ terminus was also tested (see Fig. 2). This control molecule is denoted 8GC. The DNA oligonucleotides were synthesized using an ABI 392 Synthesizer, purified by PAGE, and stored at −70°C in TE buffer. The prediction that each hairpin would adopt one basepaired structure was tested and confirmed using the DNA *m*fold server (http://bioinfo.math.rpi.edu/mfold/dna/form1.cgi), which is based in part on data from SantaLucia (1998).

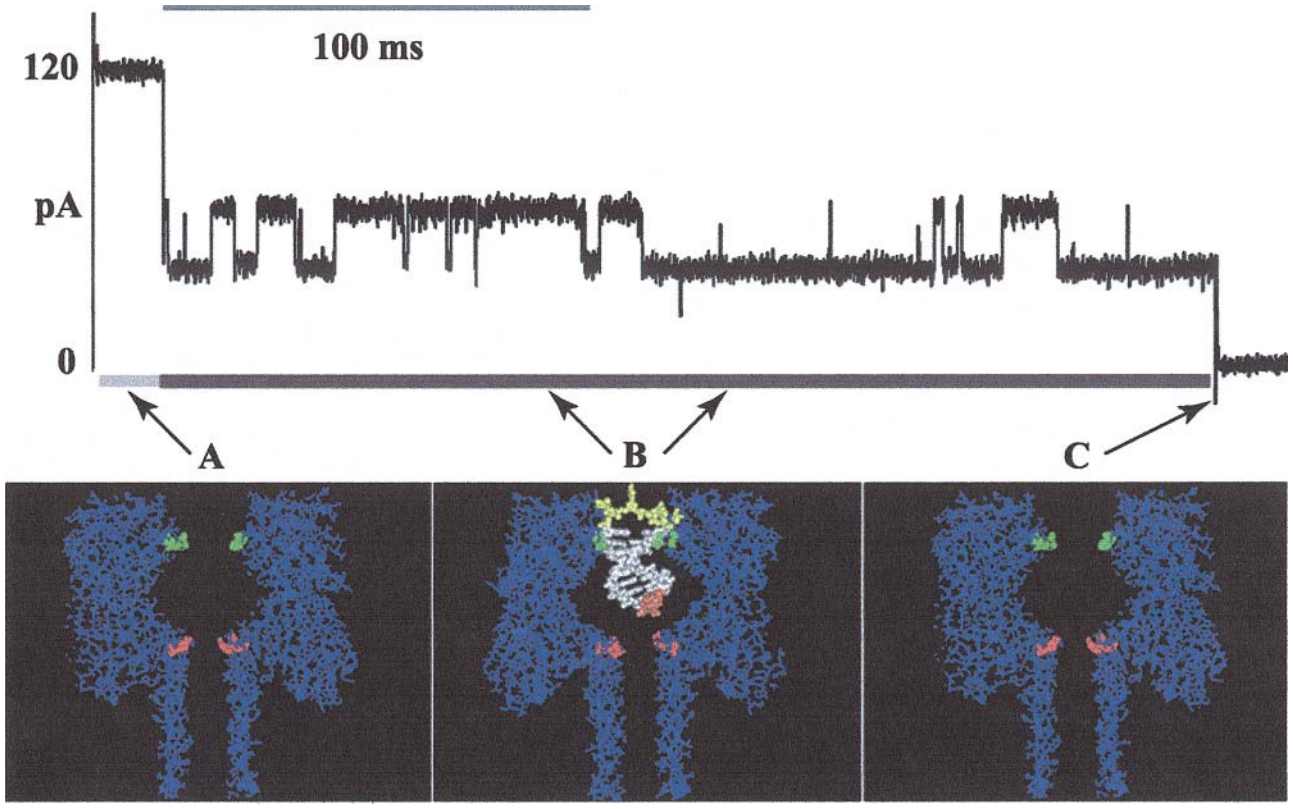

FIGURE 1 Examination of DNA duplex ends using a voltage-pulse routine. An observation cycle for a 9GC hairpin blockade event is shown. At the start of each voltage cycle the voltage across the pore is reset to 0 mV. A potential difference of 120 mV (*trans* side positive) is then applied for 250 ms, initially resulting in an open channel current of 120 pA (*image A*, with *arrow* indicating the open channel region of the current trace). In time, duplex DNA is pulled into the pore by the applied potential causing an abrupt current decrease (*image B*, with *arrows* and *solid bar* delineating region of blockade signal). After the 250-ms forward bias, the potential is briefly reversed (−40 mV, *trans* side) then set at 0 mV for 50 ms which clears the pore (*image C*, with *arrow* indicating the voltage reversal spike). The cycle is then repeated to examine the next molecule. Only the first 100 ms of blockade signal is used to identify each current signature. In the diagrams, the stick figure in blue is a two-dimensional section of the α-hemolysin pore derived from x-ray crystallographic data (Song et al.). A ring of lysines that circumscribe a 1.5-nm-limiting aperture of the channel pore is highlighted in red. A ring of threonines that circumscribe the narrowest, 2.3-nm-diameter section of the pore mouth is highlighted in green. In our working model, the four dT hairpin loop (*yellow*) is perched on this narrow ring of threonines, suspending the duplex stem in the pore vestibule (Vercoutere et al., 2001, Winters-Hilt et al., in preparation). The terminal basepair (*brown*) dangles near the limiting aperture. The structure of the 9bp hairpin shown here was rendered to scale using WebLab ViewerPro. See the Discussion section for further details on the mechanism behind the blockade signals.

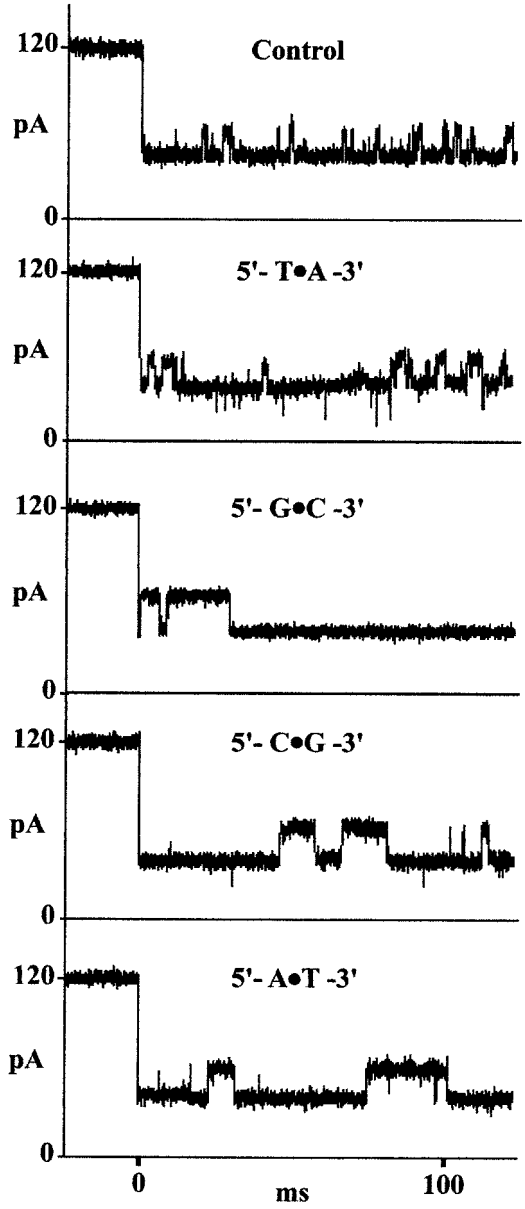

FIGURE 2 Typical blockade signatures for each of the five classes of DNA hairpins. The nine basepair hairpins differ in only their terminal basepairs. The variants were chosen to include the two possible Watson-Crick basepairs and the two possible orientations of those basepairs at the duplex ends. The core 8bp stem and 4dT loop were identical with the primary sequence 5′-TTCGAACGTTTTCGTTCGAA-3′, where the basepaired compliments are underlined. The eight basepair hairpin that was used as a control had the primary sequence 5′-GTCGAACGTTTTCGTTCGAC-3′.

Sampling protocol

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 voltage toggling protocol was based on a 300-ms cycle: 250 ms at ~120 mV for capture/measurement, followed by 1 ms at −40 mV for ejection, and then 49 ms at 0 mV for reset. The 300-ms voltage-toggle cycle was chosen to accommodate signal acquisition of the first 100 ms of blockade signal (as shown in Fig. 1). If less than 100 ms of blockade signal was acquired before the ejection phase of the cycle, the signal was ignored. The effective duty cycle for the 100-ms blockade measurements was one reading every 0.4 s.

Signal acquisition

Ionic current was filtered at 10 kHz bandwidth using an analog low pass Bessel filter and recorded at 20 μs intervals using an Axopatch 200B amplifier (Axon Instruments, Foster City, CA) coupled to an Axon Digidata 1200 digitizer. 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, Fig. 3). Two states, sequentially connected, were used for resetting and initializing the FSA. Transition between the two states, from reset-start to reset-ready, was accomplished upon measuring a short section of acceptable baseline current (200 μs). An abrupt drop in current to 70% residual current, or less, then triggered transition from the reset-ready state to the signal-active state. From the signal-active state, processing advanced to one of two states (good- and bad-end-level states) according to an end-of-signal profile. The profile rule simply required that the last end-level-range observations had to have current above minimum-end-level-value. Satisfying the rule led to the good-end-level state, otherwise the bad-end-level state was reached. If there was a normal return to baseline (good-end-level state), or a signal-blockade scan exited due to truncation (bad-end-level state), the signal complete state was reached, otherwise further scanning was performed. Further scanning involved transition through the internal active state, where local signal properties, observation less than maximum-cutoff and observation greater than minimum-cutoff, were used to decide whether to exit (to the reset-end state) or continue the blockade scan (return to the signal-active state). Similar to the local blockade signal properties that determined how to transition from the internal-active state, transition to the acquire-signal state from the signal-complete state was based on several global properties of the signal trace: maximum blockade sample less than maximum-cutoff and greater than min-max-internal, minimum blockade sample greater than minimum-cutoff and less than max-min-internal, and signal duration greater than or equal to minimum-duration.

The time-domain FSA was tuned so that it would rarely miss signal acquisitions (low false negatives) by allowing for large numbers of mistaken signal acquisitions (i.e., large false positives). The acquisition bias was accomplished by imposing constraints on valid starts that were weak while maintaining constraints on valid interior and ends that were strong. The bias toward high sensitivity permitted tuning on FSA parameters with a simplified objective. For the blockade signatures studied, the FSA parameters for maximal signal acquisition shared a broad, common range, allowing one set of FSA parameters (a single generic FSA) to acquire all signals. After tuning, the FSA parameters were: minimum-start-drop = 70%, maximum-cutoff = 170%, minimum-cutoff = −60%, end-level-value = 95%, max-min-internal = 55%, min-max-internal = 70%, end-level-range = 10 (at the 20 μs sampling this leads to a minimum 200 μs interval between blockade acquisitions), and maximum-duration = 100 ms = minimum-duration (for 100-ms truncation on acquired signals). Parameters expressed in terms of percentages refer to current measurements normalized with respect to the average baseline (determined by a 1024-element baseline sampling on the contiguous baseline segment nearest and before the blockade signal).

Signal preprocessing

Each 100-ms signal acquired by the time-domain FSA consisted of a sequence of 5000 subblockade 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 required construction of a dynamic programming table with only 625 columns). The signal preprocessing makes use of an off-line wavelet stationarity analysis (Diserbo et al., 2000). The stationarity analysis (Off-line Wavelet Stationarity Analysis, Fig. 3) was based on a training set

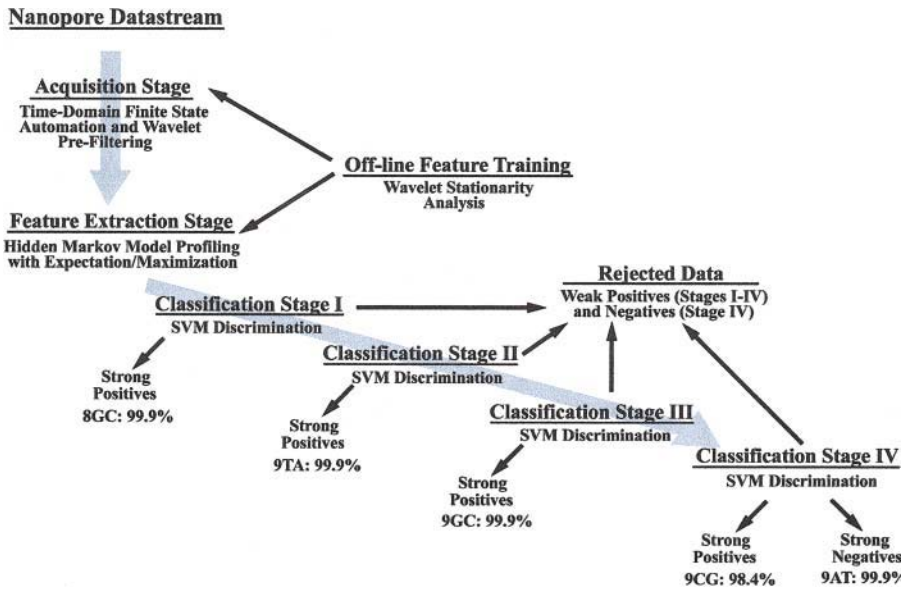

FIGURE 3 Machine learning strategy. Signal acquisition was performed using a time-domain, thresholding, Finite State Automaton. This was followed by adaptive prefiltering using a wavelet-domain Finite State Automaton. Feature extraction on those acquired channel blockades was done by Hidden Markov Model 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 inasmuch as 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 ~6 s on hardware described in Methods).

of blockade signals from each of the different classes of blockades to be discriminated.

A 1024-sample Haar wavelet transform (Nievergelt, 1999) was applied to the time-domain information at the start of each blockade in the training set. The wavelet-domain components were then completed so that a wavelet-domain FSA could easily reference a ''moving'' wavelet transform (i.e., the Haar transform with forward-shifting time-origin in the 5000-element sequence). The FSA scanning operation over wavelet components was then defined. Half the scanning data consisted of values from a 2^N-point moving average (with N equal to a specified order of wavelet component), whereas the other half of the data consisted of the order-N wavelet difference coefficients. The moving sum and difference wavelet components for a given order provided a dual track of wavelet states. For the data analyzed, the information in the difference coefficients was only used when the difference coefficient was very large, indicating a transition between blockade levels, or the start/end transition of the blockade itself. The dual track of wavelet states was thereby reduced to a single track, consisting of a sequence of sum wavelet components with the occasional occurrence of an overriding difference-wavelet component. (The single track override also provided the framework for incorporating fine-scale feature extraction, such as spike detection, from the time-domain FSA, but such feature passing was not used in the results that follow.)

A tuning process was used by the wavelet-domain FSA to select the optimal order of wavelet component to use as the basis for the signal quantization. The tuning method employed an emergent grammar heuristic on the single-track-compressed sequence of states. The method made use of the property that, as wavelet order was decreased, the difference wavelet override was triggered more easily—which eroded the distinctive transition structure seen in blockade signals. The lowest order that retained a distinctive transition structure, or grammar, was then used as the basis for the quantization. For the data examined this corresponded to $N = 3$ for an eightfold reduction in HMM processing. Blockade binning statistics on the sum wavelet components (at $N = 3$) were calculated for the different classes of channel blockade. Clearly discernable sum-wavelet characteristics were possible with a blockade state grayscale that ranged in 1% increments of the open-channel current. The 1% gray scale and $N = 3$ wavelet-order were used as the basis for state quantization by the HMM in the processing stages that follow.

Feature extraction

Hidden Markov Models (Durbin 1998) provide a statistical framework for sequences of observations obeying stationary Markov statistics. The ''hidden'' part of the HMM consists of the labelings, $\mathbf{s_i}$, for each observation, and $\mathbf{z_i}$, where the index $\mathbf{i}$ labels the observation. The stationary statistics for a first-order HMM are described in terms of emission probabilities, $\mathbf{e_{ni}} = \mathbf{p(Z_j = z_i \mid S_j = n)}$, transition probabilities, $\mathbf{a_{nm}} = \mathbf{p(S_j = m \mid S_{(j-1)} = n)}$. (The indexing on $\mathbf{j}$ is left in for clarity on the transition probability definition; from stationarity the expressions are valid for any choice of $\mathbf{j}$.) Given the above stationarity statistics, the probability for a sequence of $\mathbf{L}$ observations can be expressed as $\mathbf{p(Z_0 = z_0, \ldots, Z_{(L-1)} = z_{(L-1)})} = \Sigma_\mathbf{k} \mathbf{f_{ki} b_{ki}}$, where $\mathbf{f_{ki}}$ are the forward probabilities, $\mathbf{f_{ki} = p(Z_0 = z_0, \ldots, Z_i = z_i, S_i = k)}$, and $\mathbf{b_{ki}}$ are the backward probabilities, $\mathbf{b_{ki} = p(Z_{(i+1)} = z_{(i+1)}, \ldots, Z_{(L-1)} = z_{(L-1)} \mid S_i = k)}$. The forward and backward variable can be recursively defined by $\mathbf{f_{ki}} = \mathbf{a_{\beta k} e_{ki}\, f_{\beta(i-1)}}$ and $\mathbf{b_{ki}} = \mathbf{a_{k\beta} e_{\beta(i+1)}\, b_{\beta(i+1)}}$ (Durbin 1998), where we use the Einstein convention of implied summation over repeated Greek letter indices. The recursive definitions on forward and backward variables permit efficient computation of observed sequence probabilities using dynamic programming tables.

The recursive algorithm for the most likely state path given an observed sequence (the Viterbi algorithm) is expressed in terms of v_{ki}, the probability of the most probable path that ends with observation $Z_i = z_i$, and state $S_i = k$. The recursive relation is $v_{ki} = \max_n\{e_{ki} a_{nk} v_{n(i-1)}\}$, where the $\max_n\{\ldots\}$ operation returns the maximum value of the argument over different values of index n, and the boundary condition on the recursion is $v_{k0} = e_{k0} p_k$. The Viterbi path labelings are then recursively defined by $(S_i \mid S_{(i+1)} = n) = \text{argmax}_k\{v_{ki} a_{kn}\}$, where the $\text{argmax}_n\{\ldots\}$ operation returns the index n with maximum value of the argument. The evaluation of sequence probability (and its Viterbi labeling) take the emission and transition probabilities as a given. Estimates on the emission and transition probabilities are obtained by an Expectation/Maximization (EM) algorithm (Durbin, 1998; commonly referred to as the Baum-Welch algorithm in the context of HMMs).

An HMM 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 0–49, 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. During this estimation, the state emission distribution of each state j was constrained

to remain a Gaussian with mean $j + 20$; only the variance was adjusted. The 50 transition parameters for each state were freely readjusted, but a pseudocount of three was used to smooth the estimations. After the EM iterations, 150 parameters, described below, were extracted from the HMM. The resulting parameter vector, normalized such that (nonzero) vector components sum to unity, was used to represent the acquired signal in discrimination at the later Support Vector Machine stages.

The 150 parameters extracted from the HMM consisted of three sets of 50 parameters each. The parameters were derived from the HMM's emission and transition probabilities and the HMM's Viterbi-path statistics (Durbin 1998). In the first set, parameter λ_j, $0 <= j <= 49$, was the a posteriori estimated fraction of the time the signal was in state j, estimated using the Viterbi path (upon completion of the EM iterations). In the second set, parameter σ_j, $0 <= j <= 49$, was the variance of the Gaussian emission distribution for state j (normalized by dividing by the sum over σ_j).

To define the third parameter set (τ_j, $0 <= j <= 49$), we began with the two states in the first parameter set that had the largest locally maximal a posteriori probabilities. The posterior probability for state k was said to be *locally maximal* if it was greater than the posterior probabilities at either state $k - 1$ or state $k + 1$. The third parameter set then consisted of a weighted combination of the outgoing transition probabilities from the two states with largest locally maximal posterior probabilities. The weighting on their transition probability combination was their posterior probabilities (λ_j). This reduced a 50×50 matrix of transition parameters to 50 parameters, although preserving information about the distinctive bilevel toggling between major blockade levels that was characteristic of the data.

The normalization on each of the three sets of 50 parameters was unity before the overall feature vector normalization. Feature vector normalization then followed with division by 3. The feature vector terms thus described a (nonzero) partition of unity, a domain that was needed for SVM discrimination that used information divergences (in addition to discrimination based on the usual geometric distance measures). The parameters were nonzero due to the Bayesian origin of the probabilities. Although mixture kernels were considered over the three sets of parameters themselves (without the overall normalization), they generally did not perform as well as the best nonmixture kernels, and will not be discussed in what follows.

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 impose a training-instance void, or ''margin,'' around the decision hyperplane.

Feature vectors are denoted by x_{ik}, where index i labels the M feature vectors ($1 \le i \le M$) and index k labels the N feature vector components ($1 \le i \le N$). For the binary SVM, labeling of training data is done using label variable $y_i = \pm 1$ (with sign according to whether the training instance was from the positive or negative class). For hyperplane separability, elements of the training set must satisfy the following conditions: $w_\beta x_{i\beta} - b \ge +1$ for i such that $y_i = +1$, and $w_\beta x_{i\beta} - b \le -1$ for $y_i = -1$, for some values of the coefficients $w_1, \ldots, w_N$, and b (again using the convention of implied sum on repeated Greek indices). This can be written more concisely as: $y_i(w_\beta x_{i\beta} - b) - 1 \ge 0$. Data points that satisfy the equality in the above are known as ''support vectors'' (or ''active constraints'').

Once training is complete, discrimination is based solely on position relative to the discriminating hyperplane: $\mathbf{w}_\beta \mathbf{x}_{i\beta} - \mathbf{b} = \mathbf{0}$. The boundary hyperplanes on the two classes of data are separated by a distance $\mathbf{2/w}$, known as the ''margin,'' where $\mathbf{w}^2 = \mathbf{w}_\beta \mathbf{w}_\beta$. By increasing the margin between the separated data as much as possible, the optimal separating hyperplane is obtained. In the usual SVM formulation, the goal to maximize $\mathbf{w}^{-1}$ is restated as the goal to minimize $\mathbf{w}^2$. The Lagrangian variational formulation then selects an optimum defined at a saddle point of $\mathbf{L(w,b;}\alpha) = (\mathbf{w}_\beta \mathbf{w}_\beta)\mathbf{/2} - \alpha_\gamma \mathbf{y}_\gamma(\mathbf{w}_\beta \mathbf{x}_{\gamma\beta} - \mathbf{b}) - \alpha_0$, where $\alpha_0 = \Sigma_\gamma \alpha_\gamma$, $\alpha_\gamma \ge \mathbf{0}$ ($\mathbf{1} \le \gamma \le \mathbf{M}$). The saddle point is obtained by minimizing with respect to $\{\mathbf{w_1}, \ldots, \mathbf{w_N}, \mathbf{b}\}$ and maximizing with respect to $\{\alpha_1, \ldots, \alpha_M\}$. If $\mathbf{y_i(w_\beta x_{i\beta} - b) - 1 \ge 0}$, then maximization on α_i is achieved for $\alpha_i = \mathbf{0}$. If $\mathbf{y_i(w_\beta x_{i\beta} - b) - 1 = 0}$, then there is no constraint on α_i. If $\mathbf{y_i(w_\beta x_{i\beta} - b) - 1 < 0}$, there is a constraint violation, and $\alpha_i \to \infty$. If absolute separability is possible the last case will eventually be eliminated for all $\alpha_\mathbf{i}$, otherwise it is natural to limit the size of α_i by some constant upper bound, i.e., $\mathbf{max}(\alpha_i) = \mathbf{C}$, for all $\mathbf{i}$. This is equivalent to another set of inequality constraints with $\alpha_i \le \mathbf{C}$. Introducing sets of Lagrange multipliers, ξ_γ and μ_γ ($\mathbf{1} \le \gamma \le \mathbf{M}$), to achieve this, the Lagrangian becomes: $\mathbf{L(w,b;}\alpha,\xi,\mu) = (\mathbf{w}_\beta \mathbf{w}_\beta)\mathbf{/2} - \alpha_\gamma[\mathbf{y}_\gamma(\mathbf{w}_\beta \mathbf{x}_{\gamma\beta} - \mathbf{b}) + \xi_\gamma] + \alpha_0 + \xi_0 \mathbf{C} - \mu_\gamma \xi_\gamma$, where $\xi_0 = \Sigma_\gamma \xi_\gamma$, $\alpha_0 = \Sigma_\gamma \alpha_\gamma$, and $\alpha_\gamma \ge \mathbf{0}$ and $\xi_\gamma \ge \mathbf{0}$ ($\mathbf{1} \le \gamma \le \mathbf{M}$).

At the variational minimum on the $\{\mathbf{w}_1, \ldots, \mathbf{w}_N, \mathbf{b}\}$ variables, $\mathbf{w}_\beta = \alpha_\gamma \mathbf{y}_\gamma \mathbf{x}_{\gamma\beta}$, and the Lagrangian simplifies to: $\mathbf{L}(\alpha) = \alpha_0 - (\alpha_\delta \mathbf{y}_\delta \mathbf{x}_{\delta\beta}\, \alpha_\gamma \mathbf{y}_\gamma \mathbf{x}_{\gamma\beta})/\mathbf{2}$, with $\mathbf{0} \le \alpha_\gamma \le \mathbf{C}$ ($\mathbf{1} \le \gamma \le \mathbf{M}$) and $\alpha_\gamma \mathbf{y}_\gamma = \mathbf{0}$, where only the variations that maximize in terms of the α_γ remain (known as the Wolfe Transformation). In this form the computational task can be greatly simplified. By introducing an expression for the discriminating hyperplane: $\mathbf{f}_i = \mathbf{w}_\beta \mathbf{x}_{i\beta} - \mathbf{b} = \alpha_\gamma \mathbf{y}_\gamma \mathbf{x}_{\gamma\beta} \mathbf{x}_{i\beta} - \mathbf{b}$, the variational solution for $\mathbf{L}(\alpha)$ reduces to the following set of relations (known as the Karush-Kuhn-Tucker, or KKT, relations): *i*), $\alpha_i = \mathbf{0} \leftrightarrow \mathbf{y}_i \mathbf{f}_i \ge \mathbf{1}$, *ii*), $\mathbf{0} < \alpha_i < \mathbf{C} \leftrightarrow \mathbf{y}_i \mathbf{f}_i = \mathbf{1}$, and *iii*), $\alpha_i = \mathbf{C} \leftrightarrow \mathbf{y}_i \mathbf{f}_i \le \mathbf{1}$. When the KKT relations are satisfied for all of the α_γ (with $\alpha_\gamma \mathbf{y}_\gamma = \mathbf{0}$ maintained) the solution is achieved. (The constraint $\alpha_\gamma \mathbf{y}_\gamma = \mathbf{0}$ is satisfied for the initial choice of multipliers by setting the α-values associated with the positive training instances to $\mathbf{1/N^{(+)}}$ and the α-values associated with the negatives to $\mathbf{1/N^{(-)}}$, where $\mathbf{N^{(+)}}$ is the number of positives and $\mathbf{N^{(-)}}$ is the number of negatives.) Once the Wolfe transformation is performed it is apparent that the training data (support vectors in particular, KKT class (*ii*) above) enter into the Lagrangian solely via the inner product $\mathbf{x}_{i\beta} \mathbf{x}_{j\beta}$. Likewise, the discriminator $\mathbf{f}_i$, and KKT relations, are also dependent on the data solely via the $\mathbf{x}_{i\beta} \mathbf{x}_{j\beta}$ inner product. Generalization of the SVM formulation to data-dependent inner products other than $\mathbf{x}_{i\beta} \mathbf{x}_{j\beta}$ are possible and are usually formulated in terms of the family of symmetric positive definite functions (reproducing kernels) satisfying Mercer's conditions (Vapnik, 1999).

Binary SVMs were grouped into a classifier tree and trained to perform multiclass discrimination on five classes of DNA hairpin as shown in classification stages I–IV in Fig. 3. Tuning on the multiclass SVM architecture was done for performance optimization. Separate tuning was done on the polarization strength used in the data cleaning (see Discriminator Implementation Section). 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, $d^2(x,y) = \Sigma_k (x_k - y_k)^2$, also known as the squared l_2-norm on $(x-y)$, $[l_2(x-y)]^2 = d^2(x,y)$, is associated with the Gaussian kernel: $K_G(x,y) = \exp(-d^2(x,y)/2\sigma^2)$. In the latter case, the information divergence (relative entropy) between probability vectors x and y, $D(x\|y) = \Sigma_k x_k \log(x_k/y_k)$, can be associated with the ''Entropic kernel'': $K_E(x,y) = \exp(-[D(x\|y) + D(y\|x)]/2\sigma^2)$. The terminating SVM node of the classifier tree (stage IV in Fig. 3) 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 l_1-norm, where $l_1(x-y) = \Sigma_k |x_k - y_k|$, with kernel $K_I(x,y) = \exp(-\sqrt{l_1(x-y)}/2\sigma^2)$. 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 are trained by solving their KKT relations using the Sequential Minimal Optimization (SMO) procedure (Platt, 1998). The

method begins by selecting a pair of Lagrange multipliers, $\{\alpha_1,\alpha_2\}$, where at least one of the multipliers has a violation of its associated KKT relations (for simplicity it is assumed in what follows that the multipliers selected are those associated with the first and second training instances: $\{x_1,x_2\}$). The SMO procedure then ''freezes'' variations in all but the two selected Lagrange multipliers, permitting much of the computation to be circumvented by use of analytical reductions. By using the constraint $\alpha_\gamma y_\gamma = 0$ to eliminate references to α_1, and performing the variation on α_2, $\partial L(\alpha)/\partial\alpha_2 = 0$ leads to the following update rule: ${\alpha_2}^{\text{new}} = {\alpha_2}^{\text{old}} - y_2((f_1-y_1) - (f_2-y_2))/\eta$. Once ${\alpha_2}^{\text{new}}$ is obtained, the constraint ${\alpha_2}^{\text{new}} \le C$ must be reverified in conjunction with the $\alpha_\gamma y_\gamma = 0$ constraint. If the $L(\alpha)$ maximization leads to a ${\alpha_2}^{\text{new}}$ that grows too large, the new α_2 must be ''clipped'' to the maximum value satisfying the constraints. For example, if $y_1 \neq y_2$, then increases in α_2 are matched by increases in α_1. So, depending on whether α_2 or α_1 is nearer its maximum of C, we have $\max(\alpha_2) = \text{argmin}\{C;\ \alpha_2 + (C - \alpha_1)\}$. See (Platt 1998) for other boundary conditions and details on the b-value update.

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. The initial SVM training that resulted was interpreted in terms of the data polarization around its discriminating hyperplane, with stronger data calls defined as those further away from the hyperplane. The polarized data was separated, using a tuned cutoff, into strong positives, strong negatives, and weak signals. The SVMs were then retrained with strong positives as the new positives and the remainder (including weak positives) as new negatives. This served to shift weak positive (and negative) nondiagnostic noise to the negatives. The retrained SVMs were then biased toward use for high-confidence calling on the positives.

Testing protocol

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. Scoring was similarly possible in the context of in silico five-way mixtures (where an equal mix of the five species was considered). Scoring with comparable permutations of the train/test day separations (~80% of the days on training, 20% of the days on testing) established roughly the same performance. (Assessing the performance when training and testing are done on different days is important. When train/test data was split by random selection without regard to day of operation scoring improved greatly, but this protocol does not reflect a realistic usage scenario.) Sequential group calling was also performed, where groups (sequential packets) of blockade signals were classified as a group. The sequential group caller was based on majority-vote (with rejection on tie), and used a 10-call group size.

For true mixture test data, tens of thousands of blockade signatures were acquired, also from different days. For true 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.

Real-time operation

One of the computational goals was real-time signal calling, here taken to mean signal calling in less time than the duration of the signal itself. This goal has practical use in detector operation in that extensive data caching is not needed (detector data outflow does not exceed the throughput of the signal processing pipeline). Under the signal sampling used here (100-ms blockades acquired, 400-ms effective duty cycle) it was possible to operate signal calling ''real-time'' with an inexpensive PC (less than $1000) that had an 800 MHz Pentium III motherboard, and 512M RAM. The computer ran under Linux (a free Unix-type operating system), and used the C and Perl software packages. (The computer was part of a five-element computer network, comprised of computers with similar computational power, which was used to manage the off-line SVM training. Job control was directly managed using remote shell commands.)

RESULTS

Using the testing protocol described above, 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 achieved in less than 6 s with 99.6% accuracy. The five species of DNA hairpins consisted of a control hairpin and four hairpins that differed only in their terminal basepairs (Fig. 2). These results were for test data drawn from nanopores established on days other than those used to generate the training data. Fig. 4 shows the scoring for multiple observation days, with the number of single molecule sampling/classifications ranging from 1 to 30. At 75% weak signal rejection, ~15 classification attempts were needed to classify the type of single-species solution being sampled; final solution classification was obtained in 6 s 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. Fig. 5 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

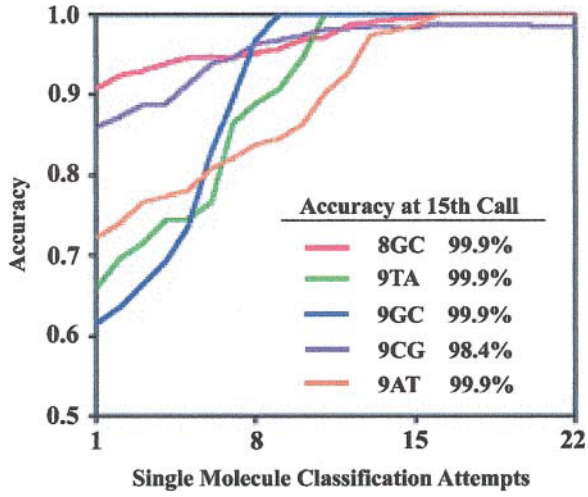

FIGURE 4 Accuracy for classification of single-species solutions of 9TA, 9GC, 9CG, 9AT, and 8GC. By the 15th classification attempt single-species solutions can be identified with high accuracy (*inset*).

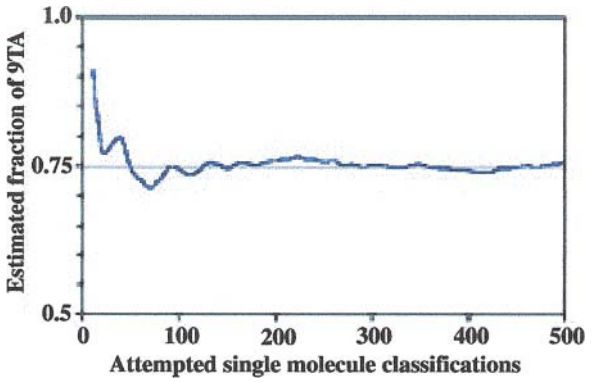

FIGURE 5 Classification on a 3:1 mixture of 9TA and 9GC hairpin molecules as a function of single molecule acquisitions. The 3:1 mol ratio is accurately identified within 1% error after 100 observations (~40 s).

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 ~100 individual molecule classification attempts (completed in 40 s).

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

The HMM-based profiling we used for feature extraction provided better discrimination than wavelet-based profiling (see Vercoutere et al., 2001). The improved signal resolution on channel blockades with HMMs is not new (Chung and Gage, 1998). (The wavelet-domain FSA that generates the blockade-level profiling does have the advantage, however, of being hundreds of times faster than the HMM processing in this instance.) The better performance with HMM processing indicated that signal analysis benefited from parsing structural information in the stochastic sequence of blockade-states. Parsing structures in stochastic data is a familiar problem in gene prediction, where Hidden Markov Models (HMMs) have been used to great advantage (Krogh et al., 1994; Stormo, 2000). Typically with gene prediction, however, HMMs are operated at a high level that parses coding starts and stops, etc., with feature scoring on starts and stops performed at a lower level by neural net or related statistical methods. For channel current analysis, the HMM extracts structural features without identifying them, effectively operating at the lower level, and used with EM (Durbin, 1998), accomplished denoising on the blockade-state structure (Chung and Gage, 1998) before extracting those features.

A single HMM/EM process was used to perform the feature extraction in our 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, 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 reduced feature set relates to physical characterizations of the captured molecule, such as level states, their time constants, and allowed level transitions.

Samples using blockade signatures of longer duration (before truncation) require fewer rejections to achieve the same signal classification accuracy. A situation that would probably favor longer signal samples than the 100 ms used here was seen in attempts to read more of the DNA hairpin end-sequence than the terminal basepair. Preliminary indications are that the penultimate basepairs can probably also be identified using longer signal samples (17 species with control). Scaling the classification task from 5 to 17 species may also require refinements to the feature extraction, such as the species-specific HMM feature extractions mentioned above.

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 (see below) 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, which 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 Fig. 3) 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.) However, if the positive and negative feature vector clusters are badly overlapped, binary SVM discrimination will be poor no matter what kernel is used. In such a circumstance, if a better choice of features cannot be obtained, rejection of low confidence data by the SVM can still be done. The SVM

confidence is a function of the distance from the feature-vector point-mapping to the separating (discriminatory) hyperplane, where greater distance represents higher confidence in discriminating between two signals.

Multiclass SVM discrimination can be obtained by grouping binary SVMs into a decision-tree architecture (Vapnik, 1999; Bredensteiner and Bennett, 1999) using rejection of low confidence data at earlier stages to postpone decisions to more appropriate later stages. All-in-one multiclass SVM optimizations are also possible (Li et al., 2001), but were not used here. Decision trees of SVMs offer good multiclass scaling properties, good noise tolerance, and low susceptibility to overtraining, but most importantly, once trained they are highly accurate and perform discriminations very quickly.

The α-hemolysin channel in the nanopore detector must be reestablished on a day-to-day basis. As a result, 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. The active stabilization methods being studied include use of reference signals (reference molecules mixed in solution) to actively track the state of the instrumentation. One stabilization approach being considered involves associative memory extensions to the feature vectors, with discrimination then operating in a higher order SVM space. Stabilization and alternative discrimination methods, such as boosting (Freund et al., 1999), could also be considered.

Blockade mechanism

Two forthcoming manuscripts (DeGuzman et al., in preparation, and Winters-Hilt et al., in preparation) will focus on details of the current blockade mechanism, so only a preliminary description is given here (Fig. 6). The intermediate level (*IL*) conductance state initiates most

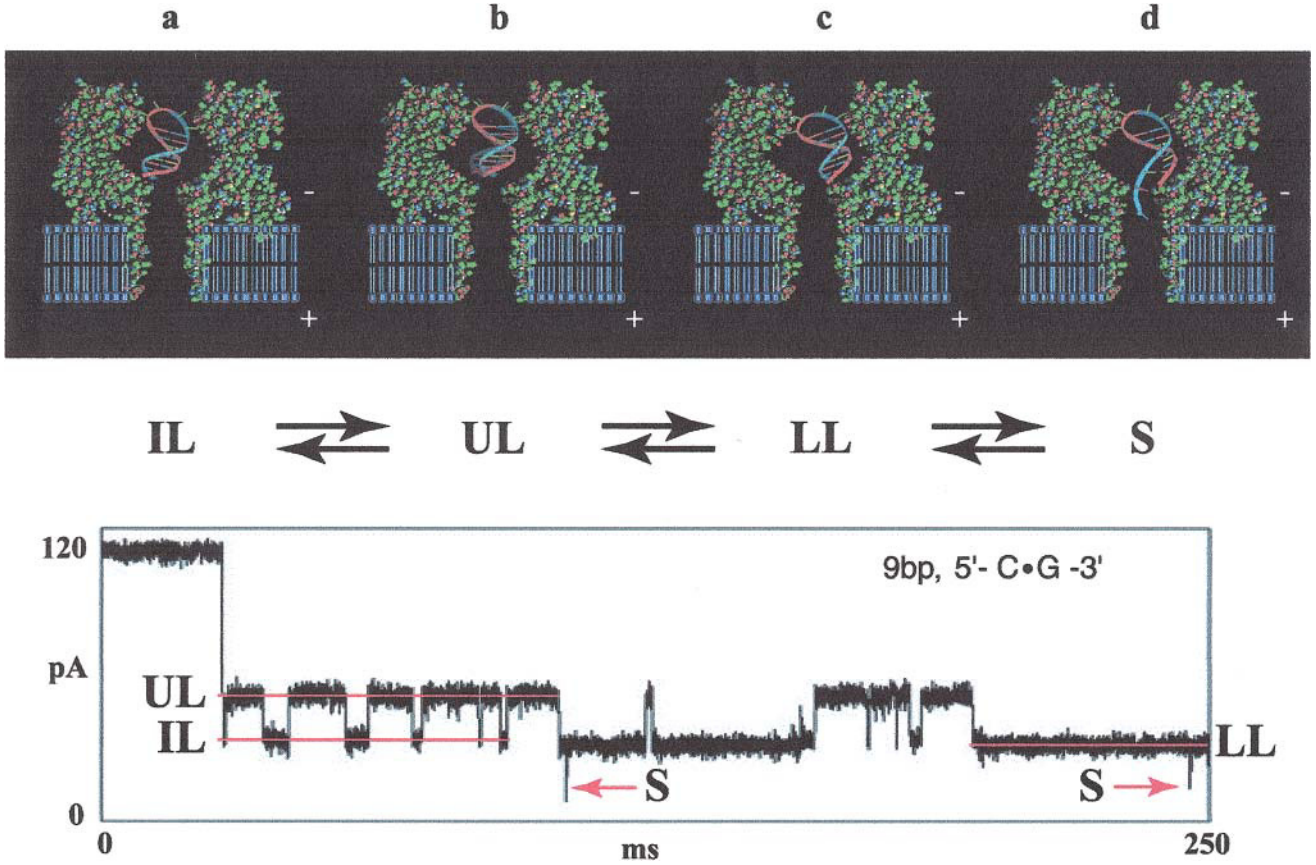

FIGURE 6 Molecular mechanisms underlying the observed current transitions. *a*) When a 9bp DNA hairpin initially enters the pore, the loop is perched in the vestibule mouth and the stem terminus binds to amino acid residues near the limiting aperture. This results in the *IL* conductance level. *b*) When the terminal basepair 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*). This *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.

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).

Applications of nanopore classification

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 confinement to reveal distinctive conductance/binding properties on those bases threading the pore's limiting aperture constriction. Single molecule classifications permit a number of technical innovations. For sequencing, the single molecule basis of measurement may permit Sanger-type sequencing on DNA molecules separated by capillary electrophoresis. If DNA can be translocated slowly enough, through a limiting aperture with dominant contributions to resistance spanning only two or three nucleotides length (~20 Ångstroms for ssDNA, 10 Ångstroms for dsDNA), then DNA sequencing of a single molecule may eventually be possible. For single nucleotide polymorphism (SNP) identification, small sample volumes can be used, such that PCR amplification may not be needed. With SNP identification, expression analysis and disease identification (for individualized therapeutics) are just a few of the possibilities. Non-PCR expression analysis may even offer a new level of experimentation on live cells using patch-clamp methods.

CONCLUSION

Five species of DNA hairpin were examined, four of which differed only in their terminal basepairs. 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 6 s and two species mixture identification in 40 s 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 retraining the machine learning components.

We thank 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, and 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.* 77:3227–3233.

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. *Computational Optimization and Applications.* 12:53–79.

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

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. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 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, editor. Academic Press, San Diego, California. 420–437.

Colquhoun, D., and F. J. Sigworth. 1995. Fitting and statistical analysis of single-channel products. *In* Single-Channel Recording, 2nd ed. B. Sakmann and E. Neher, editors. Plenum Publishing, New York. 483–587.

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

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.* 99:137–141.

Durbin, R. 1998. Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge University Press, Cambridge, England, and New York.

Freund, Y., R. Schapire, and N. Abe. 1999. A short introduction to boosting. *J. Jpn. Soc. Artif. Intell.* 14:771–780.

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. 1999. Exploiting generative models in discriminative classifiers. *In* Advances in Neural Processing Systems 11. MIT Press, Cambridge, Massachusetts.

Jelinek, F. 1997. Statistical methods for speech recognition. MIT Press, Cambridge, Massachusetts.

Joachims, T. 1998. Chapter 11: 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, Massachusetts.

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.* 93:13770–13773.

Krogh, A., I. S. Mian, and D. Haussler. 1994. A hidden Markov model that finds genes in *E. coli* DNA. *Nucleic Acids Res.* 22:4768–4778.

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.* 97:1079–1084.

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

Nievergelt, Y. 1999. Wavelets Made Easy. Birkhauser, Boston, Massachusetts.

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. 276–285.

Platt, J. C. 1998. Chapter 12: 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, Massachusetts. Ch. 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., M. R. Hobaugh, C. Shustak, S. Cheley, H. Bayley, and J. E. Gouaux. 1996. Structure of staphylococcal alpha-hemolysin, a heptameric transmembrane pore. *Science.* 274:1859–1866.

Stormo, G. D. 2000. Gene-finding approaches for eukaryotes. *Genome Res.* 10:394–397.

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.* 19:248–252.