WHITE PAPER
5500 Series SOLiD™ System

SOLiD™ System accuracy with the Exact Call Chemistry module

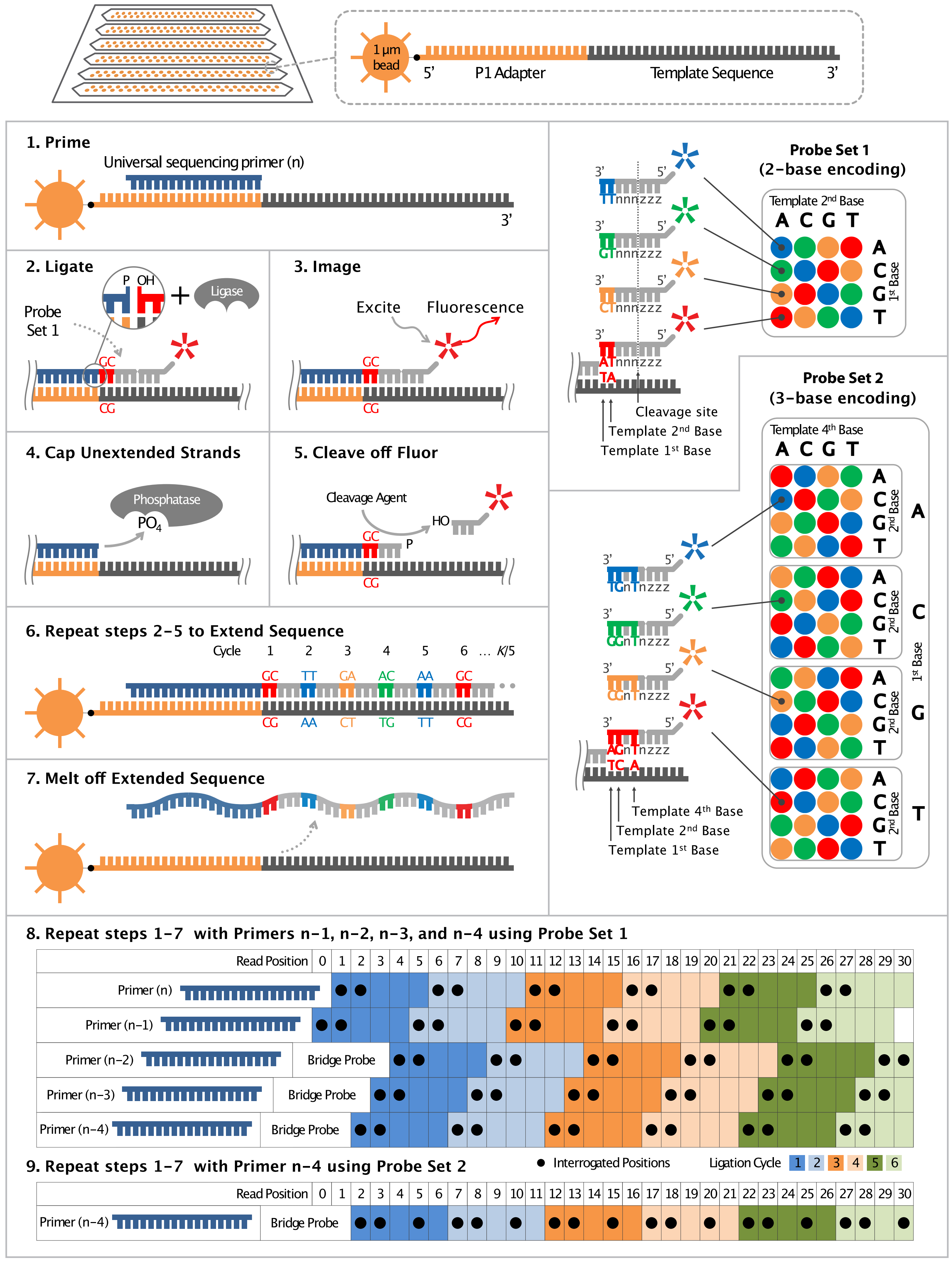
Principles of Exact Call Chemistry

Introduction

The 5500 Series SOLiD™ System is designed to provide industry-leading accuracy through a unique, ligation-based sequencing methodology, Exact Call Chemistry (ECC). This distinct sequencing approach builds on SOLiD™ 4 System chemistry, employing sequential ligation of oligonucleotide probes labeled with one of four fluorescent dyes, whereby each probe assays multiple base positions at a time. While SOLiD™ 4 System chemistry interrogates every base of the DNA template twice using two-base encoded probes, Exact Call Chemistry performs an additional inspection of the template using a new, three-base encoded probe set. This new probe set is carefully designed to complement two-base encoding and jointly form a redundant error-correction code. The set of all dye color measurements, each carrying information about multiple bases, is then used by a specialized decoding algorithm to establish the correct base sequence without knowledge of the reference, even in the presence of measurement errors. This strategy allows for error detection and correction, providing highly accurate results for resequencing, *de novo* sequencing, and rare variant detection.

Encoding of base sequences with Exact Call Chemistry

5500 Series SOLiD™ System Exact Call Chemistry is illustrated in Figure 1. Beads containing single-stranded copies of DNA library fragments are attached to the surface of the FlowChip. These sequences are interrogated by fluorescently labeled probes that hybridize and ligate at the boundary of single-stranded and double-stranded DNA. The sequencing process is organized into phases called primer rounds. Each primer round consists of hybridization of a sequencing primer with a specific offset, followed by multiple cycles of probe ligation and detection. The primer round is concluded by a reset that melts the primer and extended sequence off the template, preparing the template for the next round. SOLiD™ 4 System chemistry relies on performing five primer rounds using a two-base encoded probe set. In addition, ECC follows these five primer rounds with one additional round. This sixth round starts with the same sequencing primer as the fifth round, but uses a new, independent probe set with a specific three-base labeling. Together, the six primer rounds enable unrivaled system accuracy by forming a redundant error-correction code. The labeling of both probe sets, shown in Figure 1, is inspired by a convolutional code, a type of error-correction code used in digital communication systems where the encoded symbols are derived from a short subset of consecutive information symbols [1]. The algebraic formula used to generate them is described in Appendix C.



8. Repeat steps 1–7 with Primers n–1, n–2, n–3, and n–4 using Probe Set 1

Read Position	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
Primer (n)		●	●				●	●				●	●				●	●				●	●				●	●			
Primer (n–1)	●	●				●	●				●	●				●	●				●	●				●	●				
Primer (n–2) Bridge Probe					●	●				●	●				●	●				●	●				●	●				●	●
Primer (n–3) Bridge Probe				●	●				●	●				●	●				●	●				●	●				●	●	
Primer (n–4) Bridge Probe			●	●				●	●				●	●				●	●				●	●				●	●		

9. Repeat steps 1–7 with Primer n–4 using Probe Set 2

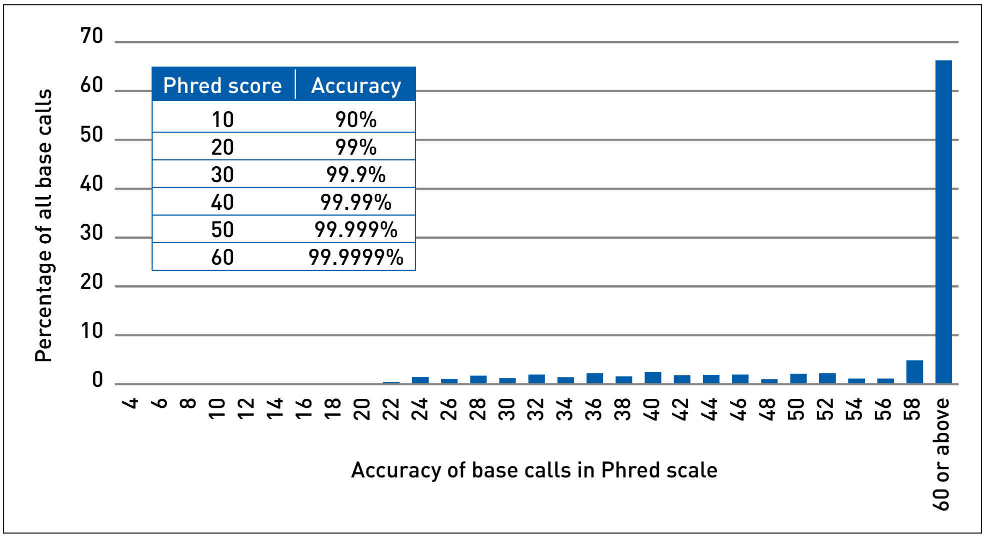
Read Position	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
Primer (n–4) Bridge Probe			●	●		●		●	●		●		●	●		●		●	●		●		●	●		●		●	●		●

Figure 1. Ligation-based sequencing with 5500 Series SOLiD™ System Exact Call Chemistry.

Demonstration of increased accuracy with Exact Call Chemistry

To validate this sequencing strategy, we performed SOLiD™ System sequencing using Exact Call Chemistry on templated beads containing synthetic DNA templates. The color measurements were collected by the instrument and processed with the ECC base calling algorithms (described in detail in Appendices A and B). Because the precise sequence of the DNA template was known, this approach specifically allowed for direct assessment of the sequencing accuracy of Exact Call Chemistry, presented in Figure 2 as a histogram of calibrated Phred scores.

As the results demonstrate, Exact Call Chemistry determines the template sequence with extremely high accuracy, the majority of base calls achieving accuracy in excess of 99.9999%. Since each base position is interrogated by multiple colors, consistent agreements between multiple color calls significantly increase the confidence in base calls. At the same time, some color errors can be corrected as a result of the single-read redundancy provided by the sixth primer round.



Phred score	Accuracy
10	90%
20	99%
30	99.9%
40	99.99%
50	99.999%
60	99.9999%

Figure 2. Histogram of base call accuracies expressed in calibrated Phred scale. Base calls were derived from color measurement through algorithms described in Appendices B and C.

Conclusion

Sequencing with Exact Call Chemistry and the 5500 Series SOLiD™ System clearly demonstrates increased performance and accuracy, which is imperative to high-throughput detection of rare genetic variants in heterogeneous or pooled samples, as well as *de novo* sequencing. The multibase encoding functionality contributes to lower error rate and reduced systematic noise, resulting in unsurpassed sequencing accuracy even at low coverage.

Bioinformatics of Exact Call Chemistry

Appendix A: Determination of base calls and quality values

Symbol	Definition
$\mathbf{u} = [u_0, u_1, u_2, \ldots, u_K]$	Sequence of bases in a DNA fragment. Base u_0 is the last base of the adapter ligated to the 5' end of the fragment, which is known ahead of time.
$\mathbf{x}_A = [x_{A,1}, x_{A,2}, \ldots, x_{A,K}]$	Expected sequence of colors resulting from interrogating **u** using probe set 1, assuming no color errors.
$\mathbf{x}_B = [x_{B,1}, x_{B,2}, \ldots, x_{B,K/5}]$	Expected sequence of colors resulting from interrogating **u** using probe set 2, assuming no color errors.
$\mathbf{y}_A = [y_{A,1}, y_{A,2}, \ldots, y_{A,K}]$	Noisy color measurements of $\mathbf{x}_A$.
$\mathbf{y}_B = [y_{B,1}, y_{B,2}, \ldots, y_{B,K/5}]$	Noisy color measurements of $\mathbf{x}_B$.

Table 1. Mathematical notation for ligation-based sequencing. Symbols and corresponding definitions are indicated.

The 5500 Series SOLiD™ System sequencing process can be understood as a transformation of the template base sequence **u** into a collection of color measurements (see Table 1 for notation). If the colors always could be read out without errors, this conversion would be a deterministic, injective function $\mathbf{u} \rightarrow (\mathbf{x}_A,\mathbf{x}_B)$, where every possible base sequence translates to a unique color sequence. For example, a sequence **u** = [TCGTGTGCTTCCGAAG] (last base T of the adapter followed by 15 template bases) is expected to translate into the set of colors in Figure 3.

In the example shown in Figure 3, 15 unknown template bases (excluding u_0) have been converted into 18 color reads. In general, there are 4^{18} (~64 billion) possible color sequences of length 18, but only 4^{15} (~1 billion) possible base sequences of length 15. This means that only one in every 64 possible color sequences corresponds to a base sequence. Such "valid" color sequences are rare and, due to the design of the probe sets, dissimilar from each other. This makes them easy to distinguish, even in the presence of some measurement noise and errors in color calls.

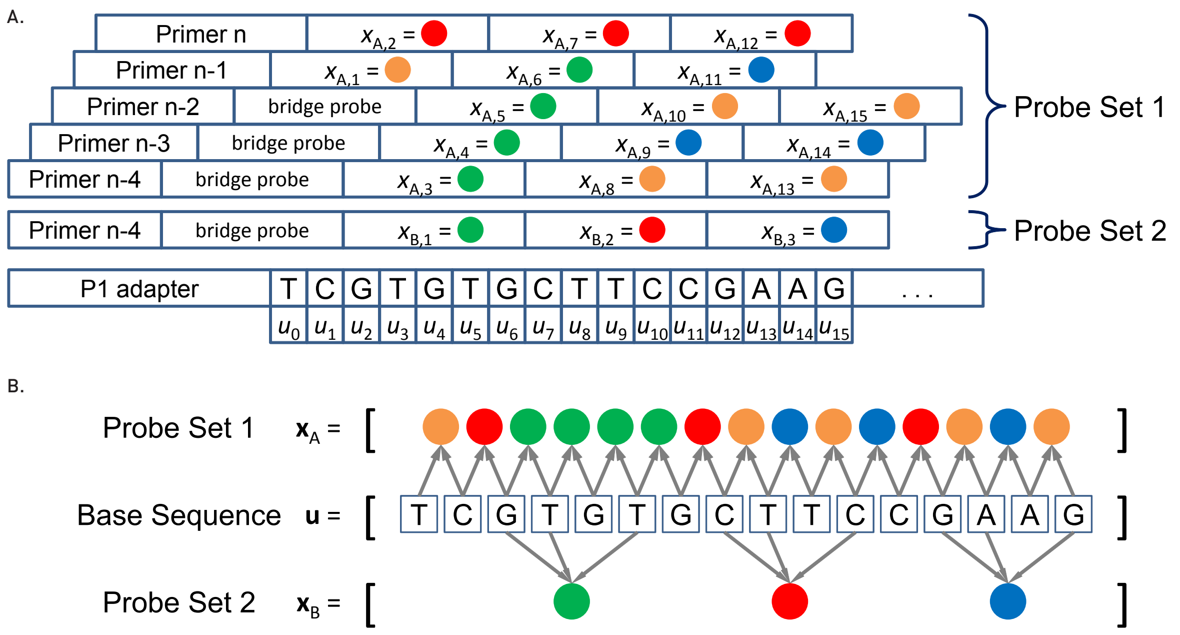


Figure 3. Example of color encoding for base sequence u = [TCGTGTGCTTCCGAAG]. Colors are arranged by (A) order of data acquisition; (B) base position and probe set.

The optics, imaging, and image processing subsystems of the 5500 Series SOLiD™ System produce color measurements for each DNA fragment and each probe ligation cycle. Measurements for a specific bead ($\mathbf{y}_A$ and $\mathbf{y}_B$, see Table 1) carry information about the color sequence $\mathbf{x}_A$, $\mathbf{x}_B$, and indirectly about base sequence **u**. The unknown sequence **u**, hidden variables $\mathbf{x}_A$ and $\mathbf{x}_B$, and observed variables $\mathbf{y}_A$ and $\mathbf{y}_B$ are linked together through a causal, statistical relation that is key to base calling (illustrated by a Bayesian network in Figure 4).

The 5500 Series SOLiD™ Sequencer performs three steps to determine individual bases and assign quality values to the calls, as illustrated in Figure 5. At the core of this process is the Bayesian inference operation that derives four conditional (posterior) probabilities, $P(u_i = \text{A,C,G,T}|\mathbf{y}_A,\mathbf{y}_B)$, for each base position *i*. The general formula for such derivation is a marginalization of $P(\mathbf{u}|\mathbf{y}_A,\mathbf{y}_B)$ derived through the Bayes theorem from conditional probability $P(\mathbf{y}_A,\mathbf{y}_B|\mathbf{u})$:

$$P(u_i|\mathbf{y}_A,\mathbf{y}_B) = \sum_{u_1}\cdots\sum_{u_{i-1}}\sum_{u_{i+1}}\cdots\sum_{u_K}\frac{P(\mathbf{y}_A,\mathbf{y}_B|\mathbf{u})\cdot P(\mathbf{u})}{P(\mathbf{y}_A,\mathbf{y}_B)}$$

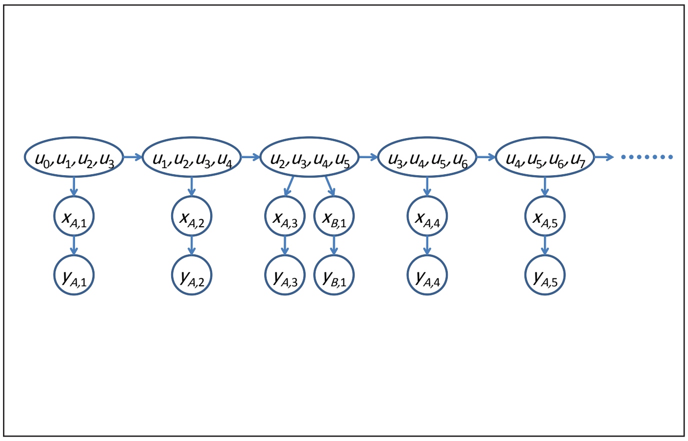


Figure 4. Bayesian network linking the unknown bases and the observed color measurements.

A practical way to efficiently perform the above computation is to use dynamic programming. The 5500 Series SOLiD™ System uses a type of dynamic programming called the forward-backward algorithm [1], which is detailed in Appendix B.

The steps preceding and following the Bayesian inference are much simpler. The measurement model, evaluated before Bayesian inference, converts each color measurement y_i into four individual color likelihoods $P(y_i|x_i =$ ●, ●, ●, ●$)$, which are later used within the forward-backward algorithm to calculate $P(\mathbf{y}_A,\mathbf{y}_B|\mathbf{u})$. Distributions $P(y_i|x_i)$ are well characterized within the 5500 Series SOLiD™ Sequencer image processing system. Finally, the product of Bayesian inference, probabilities $P(u_i = \mathrm{A,C,G,T}|\mathbf{y}_A,\mathbf{y}_B)$, are converted to base calls a_i through straightforward probability maximization:

$$a_i = \underset{u_i \in \{\mathrm{A,C,G,T}\}}{\arg\max}\; P(u_i|\mathbf{y}_\mathrm{A},\mathbf{y}_\mathrm{B})$$

Once the base call is established, its quality value b_i is directly derived from the base probability as:

$$b_i = -10\log_{10} P(u_i = a_i|\mathbf{y}_\mathrm{A},\mathbf{y}_\mathrm{B})$$

The quality values are determined from a lookup table using the four base likelihoods as feature vectors.

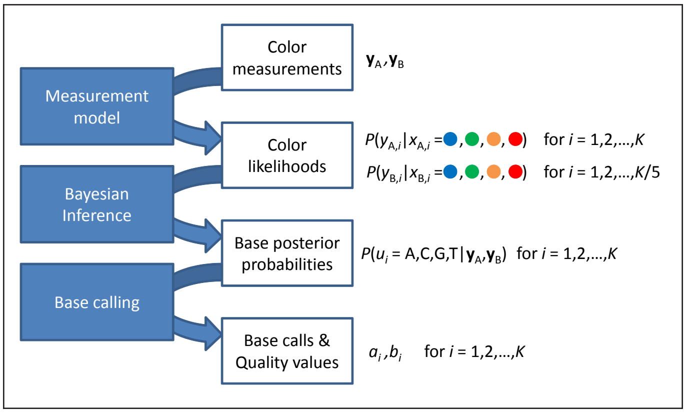

Measurement model
Bayesian Inference
Base calling
Color measurements
Color likelihoods
Base posterior probabilities
Base calls & Quality values
$\mathbf{y}_A,\mathbf{y}_B$
$P(y_{A,i}|x_{A,i} =$ ●,●,●,●$)$ for $i = 1,2,\ldots,K$
$P(y_{B,i}|x_{B,i} =$ ●,●,●,●$)$ for $i = 1,2,\ldots,K/5$
$P(u_i = \mathrm{A,C,G,T}|\mathbf{y}_A,\mathbf{y}_B)$ for $i = 1,2,\ldots,K$
a_i, b_i for $i = 1,2,\ldots,K$

Figure 5. Base calling process performed by the 5500 Series SOLiD™ System for Exact Call Chemistry.

Appendix B: The forward-backward algorithm

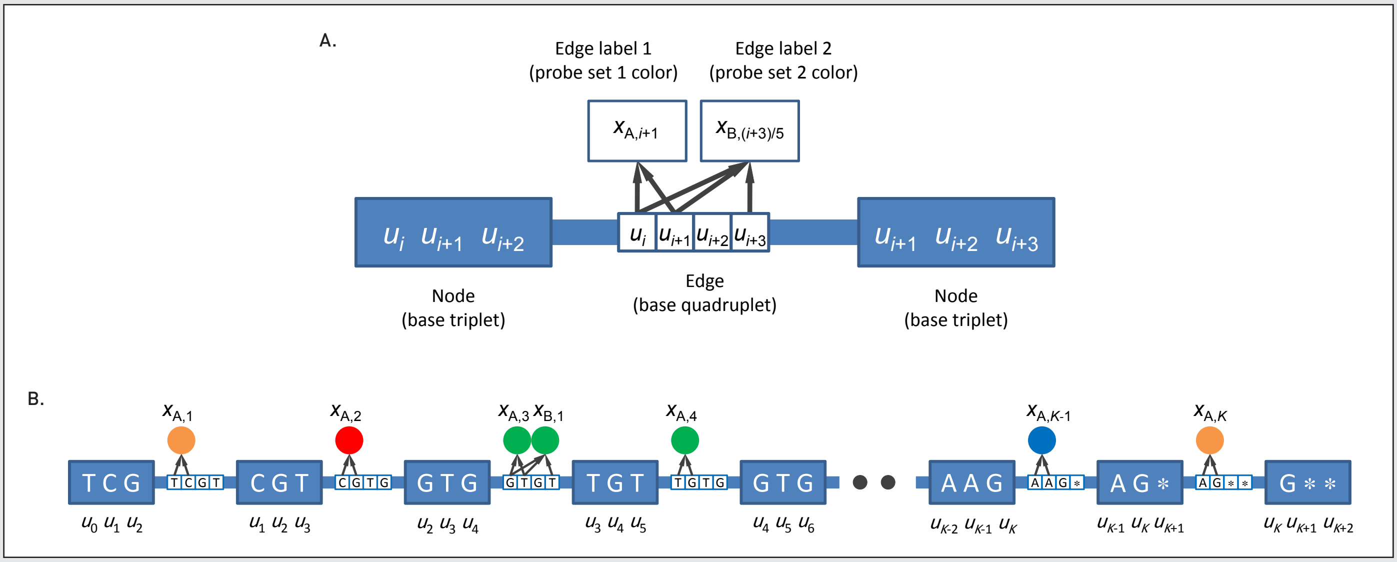


Figure 6. Graphical representation of one base sequence $\mathbf{u} = [u_0,u_1,u_2,\ldots,u_K]$ as a path through a graph. (A) Notation for variables associated with nodes and edges. (B) Example for base sequence $\mathbf{u}$ = [TCGTGTGCTTCCGAAG].

The forward-backward algorithm [1] solves the Bayesian inference problem of computing all conditional probabilities $P(u_i = \mathrm{A,C,G,T}|\mathbf{y}_A,\mathbf{y}_B)$ of bases from measurements, which is the basis for performing base calls and calculating base quality values for each base call.

The forward-backward algorithm represents any possible sequence of bases $\mathbf{u} = [u_0,u_1,u_2,\ldots,u_K]$ with a path through a graph that starts with node $[u_0,u_1,u_2]$ and passes through nodes $[u_1,u_2,u_3]$, $[u_2,u_3,u_4]$, and so on, to node $[u_{K-2},u_{K-1},u_K]$. Figure 6B shows a path corresponding to a sequence from the example in Figure 3. Every edge connecting a pair of nodes $[u_i,u_{i+1},u_{i+2}]$ and $[u_{i+1},u_{i+2},u_{i+3}]$ carries information about four bases $[u_i,u_{i+1},u_{i+2},u_{i+3}]$, which is enough to uniquely associate it with the expected color of the probes interrogating bases $[u_i,u_{i+1},u_{i+2},u_{i+3},u_{i+4}]$ from either probe set 1 $(x_{A,i+1})$ or probe set 2 $(x_{B,(i+3)/5})$ according to Figure 1. The graphical notation for values associated with nodes and edges is shown in Figure 6A.

The set of paths corresponding to all 4^K possible K-base sequences can be combined into a single compact graph called a trellis (Figure 7). The trellis guarantees one-to-one correspondence between any path from the left most nodes to the right most nodes and all possible base sequences, which makes it a blueprint for dynamic programming computations.

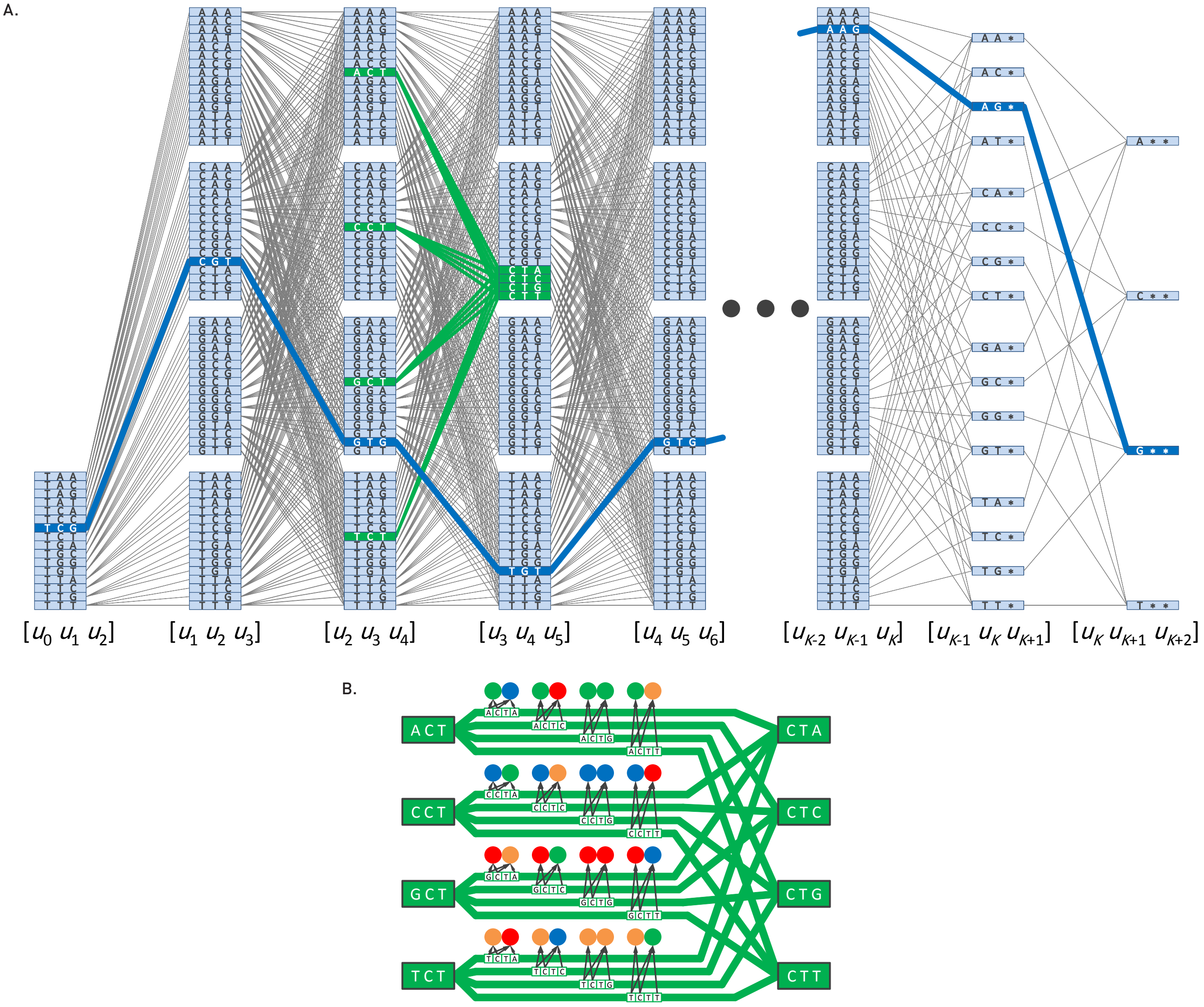


Figure 7. General structure of the trellis. [A] The trellis is divided into K stages, each consisting of 256 edges connecting two groups of 64 nodes. Each group of 64 nodes corresponds to all possible base triplets. At the beginning and end of each section there are 64 nodes, corresponding to all possible base triplets. Each edge corresponds to some quadruplet of bases [u_i, u_{i+1}, u_{i+2}, u_{i+3}], connecting node [u_i, u_{i+1}, u_{i+2}] to node [u_i, u_{i+1}, u_{i+2}, u_{i+3}]. [B] An example subset of nodes and edges. An edge that corresponds to a quadruplet of bases [u_i, u_{i+1}, u_{i+2}, u_{i+3}], has an expected color $x_{A,i+1}$ in probe set 1 and an expected color $x_{B,(i+3)/5}$ in probe set 2 determined from Figure 1. Note that measurements for probe set 2 are only available in every fifth trellis section.

The algorithm determines base calls and quality values by performing the following steps:

Step 1.

For each edge in the trellis, establish edge metric $\gamma(u_i,u_{i+1},u_{i+2},u_{i+3})$. For each edge in the graph associated with bases $u_i,u_{i+1},u_{i+2},u_{i+3}$ and colors $x_{A,i+1}$, $x_{B,(i+3)/5}$, determine the edge weight:

$$\gamma(u_i,u_{i+1},u_{i+2},u_{i+3}) = P\left(y_{A,i+1}\middle|x_{A,i+1}\right)\cdot P\left(y_{B,(i+3)/5}\middle|x_{B,(i+3)/5}\right)\cdot P(u_{i+1})$$

In the above formula, $P(y_{A,i+1}|x_{A,i+1}$ = ●,●,●,●) and $P(y_{B,(i+3)/5}|x_{B,(i+3)/5}$ = ●,●,●,●) are color likelihoods derived from color measurements $y_{A,i+1}$ and $y_{B,(i+3)/5}$. $P(u_{i+1})$ is always assumed to be 0.25. Note that for any base sequence, the product of all edge weights along the corresponding path results in $P(\mathbf{y}_A|\mathbf{x}_A) \times P(\mathbf{y}_B|\mathbf{x}_B) \times P(\mathbf{u}) = P(\mathbf{y}_A,\mathbf{y}_B,\mathbf{u})$, is proportional to the conditional sequence probability $P(\mathbf{u}|\mathbf{y}_A,\mathbf{y}_B)$. The goal of the forward-backward algorithm is the efficient marginalization of $P(\mathbf{u}|\mathbf{y}_A,\mathbf{y}_B)$ to obtain $P(u_i|\mathbf{y}_A,\mathbf{y}_B)$ for individual base positions.

Step 2.

Calculate forward node metric $\alpha(u_i,u_{i+1},u_{i+2})$ for every node in the graph:

2A. Initialize $\alpha(u_0,u_1,u_2)$ to 1 for nodes where u_0 agrees with last base of the primer and 0 otherwise.

2B. Iteratively compute all node metrics $\alpha(u_{i+1},u_{i+2},u_{i+3})$ from node metrics $\alpha(u_i,u_{i+1},u_{i+2})$ according to the formula:

$$\alpha(u_{i+1},u_{i+2},u_{i+3}) = \sum_{u_i=\mathrm{A,C,G,T}} \alpha(u_i,u_{i+1},u_{i+2})\cdot\gamma(u_i,u_{i+1},u_{i+2},u_{i+3})$$

Notice that summands in the above formula correspond to four edges arriving at the node $[u_{i+1},u_{i+2},u_{i+3}]$ from the left.

Step 3.
Calculate backward node metric $\beta(u_i,u_{i+1},u_{i+2})$ for every node in the graph:

3A. Initialize $\beta(u_K,u_{K+1},u_{K+2})$ to 1 for all nodes.

3B. Iteratively compute all node metrics $\beta(u_i,u_{i+1},u_{i+2})$ from node metrics $\beta(u_{i+1},u_{i+2},u_{i+3})$ according to the formula:

$$\beta(u_i,u_{i+1},u_{i+2}) = \sum_{u_{i+3}=\mathrm{A,C,G,T}} \beta(u_{i+1},u_{i+2},u_{i+3}) \cdot \gamma(u_i,u_{i+1},u_{i+2},u_{i+3})$$

Notice that summands in the above formula correspond to four edges arriving at the node $[u_i,u_{i+1},u_{i+2}]$ from the right.

Step 4.
Calculate partially marginalized probabilities $P(\mathbf{y}_A,\mathbf{y}_B,u_i,u_{i+1},u_{i+2})$ for all base triplets as:

$$P(\mathbf{y}_\mathrm{A},\mathbf{y}_\mathrm{B},u_i,u_{i+1},u_{i+2}) = \alpha(u_i,u_{i+1},u_{i+2}) \cdot \beta(u_i,u_{i+1},u_{i+2})$$

Each value is associated with one node in the graph.

Step 5.
Calculate the posterior base probabilities $P(u_i = \mathrm{A,C,G,T}|\mathbf{y}_A,\mathbf{y}_B)$ from partially marginalized probabilities from step 4.

$$P(u_i = \mathrm{A}|\mathbf{y}_\mathrm{A},\mathbf{y}_\mathrm{B}) = \frac{\sum_{u_{i+1}}\sum_{u_{i+2}} P(\mathbf{y}_\mathrm{A},\mathbf{y}_\mathrm{B},u_i = \mathrm{A},u_{i+1},u_{i+2})}{\sum_{u_i}\sum_{u_{i+1}}\sum_{u_{i+2}} P(\mathbf{y}_\mathrm{A},\mathbf{y}_\mathrm{B},u_i,u_{i+1},u_{i+2})}$$

All the steps of the decoding process have low complexity and are well suited for high-performance implementation.

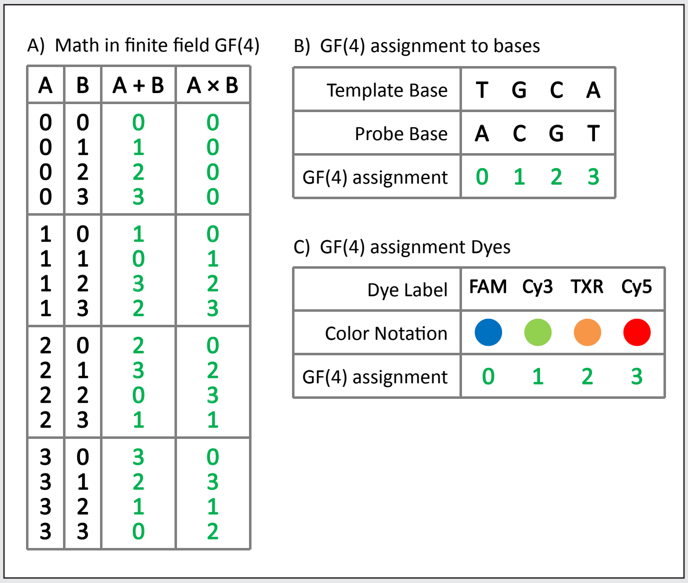
Appendix C: Probe set design

The design of the labeling for the two probe sets is inspired by a convolutional code, a type of error-correcting code used in digital communication systems [1]. Convolutional codes have two key properties. First, they have a sliding window property, where the encoded symbols are derived from a short subset of consecutive information symbols—a property that is naturally satisfied by SOLiD™ System chemistry. Second, convolutional codes are linear in a finite field; therefore, encoded symbols, when treated as elements from a finite field, can be computed as weighted sums of input symbols. This second property can be directly applied to probe set design.

Each probe consists of five nucleotides $\mathbf{v} = [v_1,v_2,v_3,v_4,v_5]$ that specifically hybridize to complementary DNA sequence, three inosines that hybridize to any sequence, and a fluorescent dye c. With four possible nucleotides (A, C, G, and T), 1,024 possible probe sequences exist, each having one of four unique dyes assigned to it. Linearity of the probe set means that each of the 1,024 probes is assigned a dye according to the formula

$$c = v_1 \times g_1 + v_2 \times g_2 + v_3 \times g_3 + v_4 \times g_4 + v_5 \times g_5,$$

where $\mathbf{g} = [g_1,g_2,g_3,g_4,g_5]$ is a vector of weights that defines the probe sets. Bases v_i, dye c, and weights g_i are considered to be elements of a finite (Galois) field of size 4, denoted GF(4) [2]. The correspondence between nucleotides v_i, dyes c and elements of GF(4) are presented in Figure 8A. Figures 8B and 8C further detail the mechanics of performing multiplication and addition of elements of GF(4). Probe sets presented in Figure 1 for Exact Call Chemistry have been generated by formula (1) by assigning weights $\mathbf{g} = [1,1,0,0,0]$ to Probe Set 1 and weights $\mathbf{g} = [1,3,0,3,0]$ to Probe Set 2.



A) Math in finite field GF(4)

A	B	A + B	A × B
0	0	0	0
0	1	1	0
0	2	2	0
0	3	3	0
1	0	1	0
1	1	0	1
1	2	3	2
1	3	2	3
2	0	2	0
2	1	3	2
2	2	0	3
2	3	1	1
3	0	3	0
3	1	2	3
3	2	1	1
3	3	0	2

B) GF(4) assignment to bases

Template Base	T	G	C	A
Probe Base	A	C	G	T
GF(4) assignment	0	1	2	3

C) GF(4) assignment Dyes

Dye Label	FAM	Cy3	TXR	Cy5
Color Notation	🔵	🟢	🟠	🔴
GF(4) assignment	0	1	2	3

Figure 8. Finite field GF(4). [A] addition and multiplication in GF(4); [B] association between elements of GF(4) and nucleotides; [C] association between elements of GF(4) and dyes.

References

1. Lin, Costello, Error Control Coding (2nd ed.), Prentice Hall, 2004, ISBN 0130426725.
2. Lang, Algebra (3rd ed.), Springer, 2002, ISBN 038795385X.

Life Technologies offers a breadth of products DNA | RNA | PROTEIN | CELL CULTURE | INSTRUMENTS

For Research Use Only. Not intended for any animal or human therapeutic or diagnostic use.

Printed in the USA. **CO18253 0311**

Headquarters
5791 Van Allen Way | Carlsbad, CA 92008 USA | Phone +1.760.603.7200 | Toll Free in the USA 800.955.6288
www.lifetechnologies.com

life technologies™