METHOD OF DNA SEQUENCE REPRESENTATION BY USING DNA SURFACE CHARGE PATTERNS AND ITS APPLICATION TO IDENTIFY DNA BINDING SITE

Field of the Invention

[0001] The present invention generally relates to DNA sequence representation and computational prediction/identification of DNA binding sites, and more particularly to a method of DNA sequence representation by using DNA surface charge patterns and its application to identify DNA binding sites.

Background of the Invention

[0002] One challenge for bioinformatics in the post-genome sequence era is to relate the DNA sequences to their functions. Fortunately, the practitioners in bioinformatics are not starting from scratch. Experiments conducted in laboratories have established that many different DNA sequences and proteins may perform identical or similar functions in the same or different organism. It is interesting to note that distinctive recurring patterns ("motif") in biological sequences such as DNA and protein have been identified for specific functions. For example, the TATA box is conserved within all identified prokaryotic promoters.

[0003] The recognition of motifs provides a promising means for computationally identifying from unknown sequences of potential DNA binding sites that contain the motif or a variant thereof. *See*, Vanet and Marsan, Promoter sequences and algorithmical methods for identifying them, Research in microbiology **150**: 779-799 (1999). For example, DNA motifs are used to ferret out DNA binding sites in unknown sequences. *See*, Stormo, G. D., DNA binding sites: representation and discovery, Bioinformatics (Oxford, England) **16**(1): 16-23 (2000).

[0004] However, the conservation of the DNA sequences of different DNA motifs varies. For example, DNA sequences for restriction enzymes are highly or strictly conserved because the restriction enzymes recognize and cut precisely at very specific DNA motif sequences. To the contrary, DNA motif sequences such as promoters and enhancers allow one or more base variations. *See*, Stormo, G. D., **16**(1): 16-23 (2000).

Thus, a DNA sequence representation that is suitable for identifying highly or strictly conserved DNA motif sequences such as restriction enzyme sites may not be useful for identification of less conserved DNA sequences such as promoters and regulatory protein binding sites. It is further appreciated that the DNA sequence representation is crucial for the improvement of identifying DNA binding sites in an unknown sequence on the basis of DNA motifs. There are a few of available DNA motif sequence representations including consensus sequences, consensus matrices, weighted matrices, and sequence logos. See, Schneider and Stephens, Sequence logos: a new way to display consensus sequences, Nucleic acids research **18**(20): 6097-6100 (1990); Schneider, T. D., Some lessons for molecular biology from information theory, Entropy measures, maximum entropy principle and emerging applications. Special series on studies in fuziness and soft computing. J. Kacprzyk. New York, Springer-Verlag. **119:** 229-237 (2003).

[0005] A consensus sequence can be defined as "the bases that appear with highest frequency at each position when a series of sequences believed to have common function are compared." *See*, Garrett and Grisham, Transcription and the regulation of gene expression, Biochemistry Florida, Saunders College Publishing (1999). It is the oldest representation used in the literature and is equivalent to a 'pattern' in the computer science community. *See*, Vanet and Marsan, **150**: 779-799 (1999); Stormo, G. D., **16**(1): 16-23 (2000). An exemplary consensus sequence is shown in FIG 1. As the simplest form of DNA motif representation, the method of consensus sequence suffers from a considerable loss of information with respect to the data set from which it is derived. For example, a consensus sequence in representing promoter sites will lose the information of the low conservation of DNA binding sites. As these functional sites are not strictly conserved, with different degrees of conservation at various base positions, a single consensus sequence is vastly inadequate in representing the allowable variations in DNA binding sites.

[0006] Using a consensus sequence to identify DNA binding sites is not very successful. By comparing the false negative and false positive rates from the consensus sequence approach, the flaws of using consensus sequences for prediction of DNA binding sites become evident. Predictions without mismatches allowed yield high specificity but extremely low sensitivity as only exact matches are found. However, increasing the

allowable mismatches greatly increases the number of false positive predictions to intolerable levels. *See*, Stormo, G. D., **16**(1): 16-23 (2000).

[0007] An improvement to the simple consensus sequence is the use of 'degenerate letters', for ambiguous nucleotides, each representing more than one nucleotide. *See*, Vanet and Marsan, **150**: 779-799 (1999); Stormo, G. D., **16**(1): 16-23 (2000). Degeneracy at certain positions in the site sequence is represented by characters specified in the IUPAC code as shown in FIG 2. *See*, Cornish-Bowden, A., Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984, Nucleic acids research **13**: 3021-3030 (1985). Although an improvement to the simple consensus sequence, the underlying nucleotide frequency information is still lost. Regular expression searches using consensus sequence with IUPAC codes yield better results than simple consensus sequences, but nevertheless fail to perform satisfactorily. *See*, Stormo, G. D., **16**(1): 16-23 (2000).

[0008] To preserve the nucleotide frequency information lost in the consensus sequence approach, consensus matrices have been developed. A consensus matrix representing a DNA binding site consists of four rows, each for one of the four nucleotides present in DNA, and as many columns as the size of the binding site it models. *See*, Vanet and Marsan, **150**: 779-799 (1999); Stormo, G. D., **16**(1): 16-23 (2000). Each cell *M(i,j)* of the matrix corresponds to the weight assigned to the particular nucleotide *i* at position *j* of the presumed binding site, where *i* is the row number and *j* the column. FIG 3a shows an exemplary alignment matrix, where each cell of the matrix corresponds to the raw frequency of each nucleotide at each position in the aligned binding sites. *See*, Hertz and Stormo, Identifying DNA and protein patterns with statistically significant alignments of multiple sequences, Bioinformatics **15**(7-8): 563-577 (1999). It is apparent that consensus matrices are probabilistic in nature. *See*, Brazma, et al., Approaches to the automatic discovery of patterns in biosequences, Journal of computational biology **5**(2): 279-305 (1998).

[0009] Another type of consensus matrices is the weighted matrix. The weighted matrix is populated by weights that are used to score a test sequence against the binding site pattern. FIG 3b shows an exemplary weighted matrix that is obtained by the transformation of the alignment matrix as shown in FIG 3a. The transformation may be done by applying the formula $\ln \frac{(ni,j + pi)/(N+1)}{pi}$ (log likelihood) to the alignment matrix,

where $n_{i,j}$ is the number of times the nucleotide letter i appears at the position j, N the total number of sequences, and p_i is the a priori probability of the nucleotide letter i (frequency of nucleotide i in the genome). *See*, Hertz and Stormo, Bioinformatics **15**(7-8): 563-577 (1999).

[0010] To measure how closely a test sequence matches the biding side, the test sequence is aligned along the weighted matrix and the weights for each nucleotide at the aligned position are added; sequences with higher weight scores are thought to match the binding site better than sequences with lower weight scores. As shown in FIG 4, the test sequence (AGGTGC) has a total score of $(1.2)+(0.96)+(0.96)+(0.59)+(0.59)=4.3$. Various other weights have been also used to define a consensus matrix, including pseudo counts and Dirichlet priors. *See*, Hertz and Stormo, Bioinformatics **15**(7-8): 563-577 (1999).

[0011] While consensus or weighted matrices provide a better representation than does the consensus sequence, the results of DNA binding site searches using weight matrices remain unsatisfactory. Possible reasons for the disappointing results may include the assumption that the bases within the DNA binding site is of independence to each other. *See*, Fickett, J. W., The gene identification problem: An overview for developers, Computers & chemistry **20**(1): 103-118 (1996). In addition, nucleotides in promoter or DNA binding activity are thought not to be independent and their weights not be additive across the positions of the promoter. *See*, Vanet and Marsan, **150**: 779-799 (1999).

[0012] In an attempt to incorporate base frequencies into the familiar consensus sequence, Schneider, et al. developed a graphical binding site representation through the use of variable height letters to depict quantitatively the average sequence conservation, as well as the base frequencies, in a set of pre-aligned sequences. *See*, Schneider and Stephens, Sequence logos: a new way to display consensus sequences, Nucleic acids research **18**(20): 6097-6100 (1990); Schneider, T. D., Some lessons for molecular biology from information theory, Entropy measures, maximum entropy principle and emerging applications. Special series on studies in fuziness and soft computing. J. Kacprzyk. New York, Springer-Verlag. **119:** 229-237 (2003). While essentially a one-dimensional method, the approach may shed some light on the three dimensional interactions.

[0013] Information Theory was used to calculate information in binding sites. Schneider defines the information R at each site in terms of the variation in uncertainty H,

equation (1). *See*, Schneider, T. D., Entropy measures, maximum entropy principle and emerging applications. Special series on studies in fuziness and soft computing. J. Kacprzyk. New York, Springer-Verlag. **119:** 229-237 (2003). H_{before} is taken to be 2 assuming that each of the four nucleotides occur with the same frequency of 0.25, equation (2). Equation (3) shows the derivation for H_{after}.

[0014] $R = H_{before} - H_{after}$ (1)

[0015] $H_{before} \cong 2$ (bits/base) (2)

[0016] $H_{after} = -\sum_{b} f_{(b,l)} \log_2 f_{(b,l)}$ (3)

[0017] where $f_{(b,l)}$ = number of occurrences of base b in position l divided by the number of binding sites. In a highly conserved site, the value returned by equation (3) will be nearly zero, and the change in information shown will be around 2 full bits.

[0018] Schneider showed the presence of two peaks in most binding sites that are about 10 base pairs apart. These peaks corresponds to one DNA helix turn, as well as the locations of two consecutive major grooves on the DNA double helix. A detailed diagram of a sequence logo together with its derivative sequences and explanatory notes is shown in FIG 5. Although the sequence logo improves the DNA representation over the consensus sequences, it suffers the same problem of rigidity of base positions.

[0019] In short, the existing DNA sequence representations represent DNA sequences at 'mono-nucleotide' level, i.e. representing nucleotide base pairs within any DNA sequences as alphabets (consensus sequence), numerical values (consensus and weighted matrices), or graphical representations (sequence logos). By virtue of their reliance on representing DNA sequences at individual base levels, DNA binding sites prediction programs that employs existing DNA sequence representations have yielded disappointing results. For example, current promoter prediction programs based on the aforementioned DNA sequence representations generates about 30% - 40% false negative predictions and about 45% - 60% false positive predictions for human genomes. *See,* Duret and Bucher, Searching for regulatory elements in human noncoding sequences, Current opinion in structural biology 7(3): 399-406 (1997). The performances of prokaryotic promoter prediction algorithms using the same DNA sequence representations are equally poor. *See*, Vanet and Marsan, **150**: 779-799 (1999).

[0020] Therefore, there is an imperative need to develop a new method for DNA sequence representation so as to achieve better computational prediction/identification of functional variants to a known sequence. This invention satisfies this need by disclosing a method of DNA sequence representation by using DNA surface charge patterns and its applications to identify DNA functional variants. Other advantages of this invention will be apparent with reference to the detailed description.

Summary of the Invention

[0021] The present invention provides a method of DNA sequence representation by using surface charge patterns. In one embodiment of the present invention, the method comprises the following steps of obtaining a DNA sequence represented by linear nucleotide bases, N_1 N_2 N_3 N_4 ...N_{x-1} N_x, x = the total number of the nucleotide bases of the DNA sequence; wherein N is any nucleotide base selected from the group of A, T, G, and C; adding 'di-nucleotides' into every inter-nucleotide space so that the DNA sequence may be represented by N_1 N_1-N_2 N_2 N_2-N_3 N_3 N_3-N_4 N_4 ...N_{x-1} N_{x-1}-N_x N_x, wherein N_{x-1}-N_x is any 'di-nucleotides' selected from the group of AA, AC, AG, AT, CA, CC, CG, CT, GA, GC, GG, GT, TA, TC, TG and TT; and substituting surface charge patterns of each nucleotide base and 'di-nucleotides' with the DNA sequence N_1 N_1-N_2 N_2 N_2-N_3 N_3 N_3-N_4 N_4 ...N_{x-1} N_{x-1}-N_x N_x; thereby the DNA sequence is represented by a sequence of surface charge patterns.

[0022] The present invention also provides a DNA sequence represented by a sequence of surface charge patterns. In one embodiment of the present invention, the DNA charge sequence includes surface charge patterns of each single nucleotide base and 'di-nucleotides'; wherein the each single nucleotide base is selected from the group of A, T, G, and C; wherein the 'di-nucleotides' is selected from the group of AA, AC, AG, AT, CA, CC, CG, CT, GA, GC, GG, GT, TA, TC, TG and TT; and wherein the 'di-nucleotides' are inserted into the inter-space between two adjacent nucleotide base and consist of the two nucleotide bases abutting the insertion point.

[0023] The present invention further provides a method for identifying a DNA binding site by using the DNA sequences represented by a sequence of surface charge patterns. In one embodiment of the present invention, the identification method comprises the following steps of aligning at least one known DNA binding site sequence, wherein the

at least one known DNA binding site sequence is represented by a sequence of surface charge patterns including surface charge patterns of each single nucleotide base and 'di-nucleotides'; wherein the each single nucleotide base is selected from the group of A, T, G, and C; wherein the 'di-nucleotides' is selected from the group of AA, AC, AG, AT, CA, CC, CG, CT, GA, GC, GG, GT, TA, TC, TG and TT; and wherein the 'di-nucleotides' are inserted into the inter-space between two adjacent nucleotide base and consist of the two nucleotide bases abutting the insertion point; deducing a consensus representation of the aligned at least one DNA binding site sequence; and comparing an unknown DNA sequence with the consensus representation; thereby the closer the unknown DNA sequence resembles the consensus representation, the more likely the unknown DNA sequence is a potential DNA binding site.

[0024] The objectives and advantages of the invention will become apparent from the following detailed description of preferred embodiments thereof in connection with the accompanying drawings.

Brief Description of the Drawings

[0025] Preferred embodiments according to the present invention will now be described with reference to the Figures, in which like reference numerals denote like elements.

[0026] FIG 1 illustrates an exemplary consensus sequence and its derivation from the alignment of multiple sequences.

[0027] FIG 2 is the IUPAC code table using specified characters to represent the degeneracy.

[0028] FIG 3a shows an exemplary alignment matrix, where each cell of the matrix corresponds to the raw frequency of each nucleotide at each position in the aligned DNA binding sites.

[0029] FIG 3b shows an exemplary weighted matrix that is obtained by the transformation of the alignment matrix as shown in FIG 3a.

[0030] FIG 4 shows the calculation of match scores of the Test Sequence using a weighted matrix from FIG 3b.

[0031] FIG 5 shows a detailed diagram of a sequence logo together with its derivative sequences and explanatory notes.

[0032] FIGS 6a-d show surface charge views at 'mono-nucleotide' level.

[0033] FIG 7 shows the generation of 'mono-nucleotide' level surface charge representations.

[0034] FIGS 8a-d show surface charge views at 'di-nucleotide' level.

[0035] FIG 9 shows the generation of 'di-nucleotide level' surface charge representations.

[0036] FIG 10 shows six (6) ways that DNA bending occurs.

[0037] FIG 11 shows the transformation of a DNA sequence into a surface charge representation.

[0038] FIG 12 shows the computation of scores using surface charge representation.

[0039] FIG 13 shows results of jack knife tests for "Octamer-Binding Transcription Factor 1" using the surface charge representation.

[0040] FIG 14 shows results of jack knife tests for "CAAT/Enhancer Binding Protein Beta" using the surface charge representation.

Detailed Description of the Invention

[0041] The present invention may be understood more readily by reference to the following detailed description of certain embodiments of the invention.

[0042] Throughout this application, where publications are referenced, the disclosures of these publications are hereby incorporated by reference, in their entireties, into this application in order to more fully describe the state of art to which this invention pertains.

[0043] DNA-protein interactions are three dimensional in nature. The binding of proteins to DNA requires sequence-specific DNA binding sites and sequence-specific proteins in specific secondary and tertiary structural conformations, and may even require DNA distortions to allow close interaction between the molecules of the proteins and DNA binding sites. *See*, Branden and Tooze, Introduction to protein structure. New York, Garland Pub (1999). As defined herein, the "DNA binding site" refers to a DNA sequence that can be specifically recognized and bound by a protein, a peptide or the like.

[0044] As discussed above, existing representations for DNA binding site are based solely on the order of the nucleotides found in DNA sequences. It is apparent that these

methods are an over-simplification of a complex three-dimensional problem. It is well known that most protein-DNA binding involves multiple interactions including hydrogen bond, van der Waals contact and water mediated bond formed between amino acid of proteins and base pairs of double helix DNA. Most interactions are observed to occur in the major groove of the B-DNA (naturally occurring double helix DNA). *See*, Branden and Tooze, Introduction to protein structure. New York, Garland Pub (1999); Luscombe, et al., Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level, Nucleic acids research **29**(13): 2860-2874 (2001). Thus, the use of charges in DNA binding site recognition and prediction seems to be more appropriate than nucleotide alphabets representation. However, the attempts that have been made so far to identify DNA binding site by using hydrogen donor / receptor sites and charges have focused on the positions of charged atoms at the 'mono-nucleotide' level, resulting limited success. *See*, Seeman, et al., Sequence-specific recognition of double helical nucleic acids by proteins, Proc Natl Acad Sci USA **73**(3): 804 – 808 (1976); Branden and Tooze, Introduction to protein structure. New York, Garland Pub (1999).

[0045] The present invention provides a method of DNA sequence representation by using surface charge patterns at 'di-nucleotide' level. While the present invention will use the identification of DNA binding sites as an exemplary application of the new DNA representation, it is to be appreciated that the applications of the DNA representation method disclosed herein include, but are not limited to, gene prediction and sequence alignment.

[0046] As an initial step to construct the surface charge patterns for DNA sequence representation, there is provided a brief description of derivation of surface charge patterns for individual nucleotides.

[0047] Most of the protein-DNA interactions are thought to occur in the major groove of the DNA double helix. *See*, Branden and Tooze, Introduction to protein structure. New York, Garland Pub (1999). To obtain the surface charge patterns of the major groove of each of the nucleotide base pair, the base pair is first rotated such that the major groove of the nucleotide base pair is maximally exposed. FIGS 6a-d show different views of an A-T base pair (designated as $^{5'}$ A $^{3'}$): a) Top view of A-T base pair; b) Front view with axis of A-T base pair rotated to yield maximum view of Major Groove; c) Surface charge view (front view) of A-T bas pair; and d) Surface charge view (front view)

with positive and negative surface charge areas highlighted. Diagrams were generated by WebLab ViewerLite 4.0.

[0048] After the surface charges of each nucleotide are viewed as FIG 6d, an array of 4 boxes ("Surface charge windows") are used to indicate the positions of surface charge areas that are thought to be important for protein-DNA interaction at the 'mono-nucleotide' level as shown in FIG 7. Each window of the Surface charge windows is examined and assigned an integer value of 1 for negatively charged areas or 0 otherwise. FIG 7 shows the surface charge view (front view) of the 4 nucleotide bases, A-T base pair ($^{5'}$A$^{3'}$), C-G base pair ($^{5'}$C$^{3'}$), G-C base pair ($^{5'}$G$^{3'}$), and T-A base pair ($^{5'}$T$^{3'}$), together with Surface charge windows and their assigned 'mono-nucleotide' level surface charge representations. Diagrams were generated by WebLab ViewerLite 4.0. The surface charge representation at the 'mono-nucleotide' level is summarized in Table 1.

[0049] Table 1. Summary of surface charge representation at the 'mono-nucleotide' level.

[0050]

Nucleotide	Surface Charge Representation (Mono-nucleotide level)
A	(1, 0, 1, 0)
C	(1, 1, 0, 0)
G	(0, 0, 1, 1)
T	(0, 1, 0, 1)

[0051] Now referring to FIGS 8a-d and 9, there is provided a brief description of the derivation of the surface charge patterns at the 'di-nucleotide' level. To determine the surface atomic charge patterns at the 'di-nucleotide' level, the DNA structure is rotated such that the spaces between two consecutive DNA base pairs are maximally exposed, which is about 17 degrees angle rotation from the plane that maximally exposes the major groove of the first nucleotide base pair as shown in FIGS 8a-d.

[0052] FIG 8a is a top view showing the rotation angle at which the major groove between two consecutive nucleotide base pairs is exposed. Line 1 represents the angle at which the top nucleotide base pair is maximally exposed, while the Line 2 represents the angle at which the bottom nucleotide base pair is maximally exposed. Line 3 represents the angle at which the space between 2 nucleotide base pairs is maximally exposed. Angle 'a' is equal to angle 'b', which is approximately 17 degrees. FIG 8b is the front view (major groove) of two nucleotide base pairs as shown in FIG 8a. FIG 8c is a surface

charge view (front view) of two nucleotide base pairs as shown in FIG 8b. FIG 8d is a surface charge view (front view) of two nucleotide base pairs as shown in FIG 8c with positive and negative surface charge areas highlighted. Diagrams were generated by WebLab ViewerLite 4.0.

[0053] Then, an array of 3 boxes ("Surface charge windows") are used to indicate the positions of surface charge areas that are thought to be important for protein-DNA interaction at the 'di-nucleotide' level as shown in FIG 9. FIG 9 shows the surface charge views (front view) of the four of the sixteen possible 'di-nucleotide' sequences, together with the assigned 'di-nucleotide' level surface charge representations. The 'di-nucleotide' level surface charge representation consist of an array of three integers, where 1 denotes that the interaction with a positively charge atom from an amino acid is possible (surface charge windows with bi-directional arrows), and where 0 denotes that the interaction with a positively charge atom from an amino acid is not possible. Diagrams were generated by WebLab ViewerLite 4.0. The width of the first and last of the 'di-nucleotide' surface charge windows are larger in comparison to the one located in the middle of the array to compensate for greater movement of charged atoms due to bend of DNA helix including twisting, sliding and sliding as shown in FIG 10. This results in 'di-nucleotide' level surface charge representation consisting of an array of 3 integers for each of the each of the 16 possible DNA 'di-nucleotide' sequences, namely AA, AC, AG, AT, CA, CC, CG, CT, GA, GC, GG, GT, TA, TC, TG and TT. The surface charge representation at the 'di-nucleotide' level is summarized in Table 2.

[0054] Table 2. Summary of surface charge representation at the 'di-nucleotide' level.

[0055]

Nucleotide Sequence	Surface charge Representation	Nucleotide Sequence	Surface charge Representation
AA	(1, 0, 1)	GA	(1, 1, 0)
AC	(0, 0, 1)	GC	(0, 1, 0)
AG	(1, 0, 0)	GG	(1, 0, 0)
AT	(1, 0, 1)	GT	(1, 0, 0)
CA	(0, 0, 1)	TA	(1, 1, 1)
CC	(0, 0, 1)	TC	(0, 1, 1)
CG	(0, 0, 0)	TG	(1, 0, 0)
CT	(0, 0, 1)	TT	(1, 0, 1)

[0056] Now a DNA sequence can be represented by its surface charge pattern based on surface charges representations at the 'mono-nucleotide' and 'di-nucleotide' levels discussed above. FIG 11 illustrates the transformation of a DNA sequence (AATT) into a surface charge representation.

[0057] Thus far, the principles of the DNA sequence representation by using surface charge patterns have been illustrated by employing integer value 1 for areas that are negatively charged and 0 otherwise. It is to be appreciated that the surface charges may be divided further into more groups, e.g., according to the charge types, and the representation should not be limited to using positive integers. Furthermore, the size, shape or position of Surface charge windows that is important for protein-DNA interaction are also variable. In addition, the assignment values to Surface charge windows should not be limited to their ability to bind to a positively charged atom of the binding protein as their ability to interact to negatively charged atoms of the binding protein could also be considered.

[0058] The new DNA sequence representation by using surface charge patterns will have many applications. As an illustration, the present invention further discloses that the surface charge representation of DNA sequences can be used to identify and predict DNA binding sites. The identification and prediction can be accomplished by many methods including consensus matrices and weighted matrices of charges, pattern recognition methods and methods of artificial Intelligence. As an illustration, the present invention provides a brief description of using a surface charge consensus matrix to identify/predict DNA binding sites.

[0059] Now referring to FIG 12, the initial step for predicting/identifying DNA binding sites by using the surface charge consensus matrix is to convert DNA sequences into surface charge representations. Then the consensus matrix is obtained by calculating of the average of each column of the array of integers. For example, as shown in FIG 12,

the matrix value of the first column of the consensus matrix (cell b7) is calculated by the average of cell b2 and cell b5.

[0060] Still referring to FIG 12, after the consensus matrix is established, the score of each column of a comparing sequence is obtained by multiplication of the comparing sequence's array cell with the corresponding cell of the consensus matrix. For example, for calculating the score of sequence seq4 at cell b23, the multiplication of cell b22 is performed with the corresponding cell of the consensus matrix (cell b7). The total score of the comparing sequence is the summation of the values in each of the score cells. For example, the Total Score for Seq4 is the summation of the values from cell b23 to ag23, which is 14.0 as shown in cell ah23. The maximum possible score for any sequence is the summation of all values in the consensus matrix. For example, the maximum possible score for any sequence is the summation of values from cell b7 to ag7, which amounts to 16.5 as shown in cell ah7 of FIG 12. The closer the score of a comparing sequence to the maximum score, the more likely the protein that binds to the sequences that constitute the consensus matrix will bind to the comparing sequence. For the sequences in FIG 12, the protein that binds to Seq1 and Seq2 will be more likely to bind Seq4 with the score of 14.0 than Seq3 with the score of 11.0.

[0061] Examples

[0062] The following non-limiting examples are provided by way of illustration of certain aspects of the present invention.

[0063] To test the effectiveness of using the surface charge pattern representation to predict/identify DNA binding sites, actual DNA sequences of known DNA binding sites for "Octamer-Binding Transcription Factor 1" (PDB id: 1GU4, 1GU5, 1GTW, 1H88) and "CAAT/Enhancer Binding Protein Beta" (PDB id: 1GT0, 1HF0_1, 1HF0_2) were obtained for RCSB Protein DataBank (http://www.rcsb.org/pdb/).

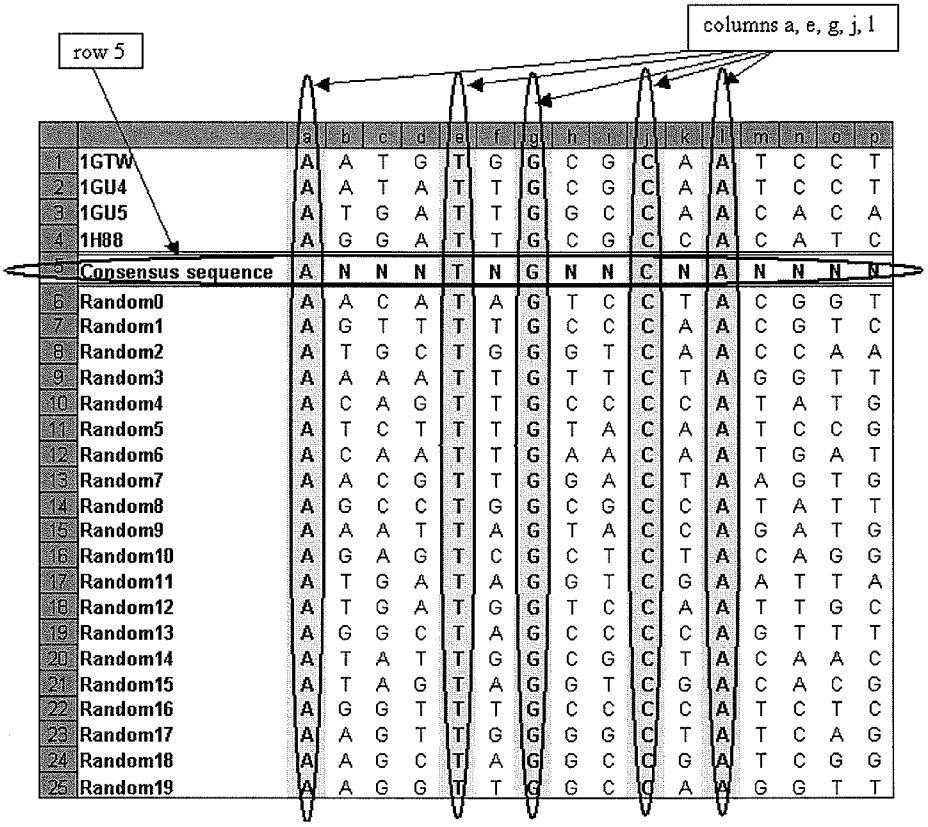
[0064] As a comparative control, a consensus sequence was first generated based on the actual DNA sequences. If there is no consensus at one position, the position is represented by 'N' in consensus sequences. Furthermore, random sequences were generated by inserting at the 'N' position a randomly selected nucleotide (A, C, T or G) with equal probabilities.

[0065] Table 3 displays the Consensus Sequence for "Octamer-Binding Transcription Factor 1" in row 4 and nineteen (19) random sequences Random[0 – 19] that

were generated by using conserved nucleotide bases in columns f, h, i, j and k, and either of the 4 possible bases in the rest of the columns, wherein each was randomly generated with equal probabilities.

[0066] Table 3. Consensus Sequence for "Octamer-Binding Transcription Factor 1"

[0067]

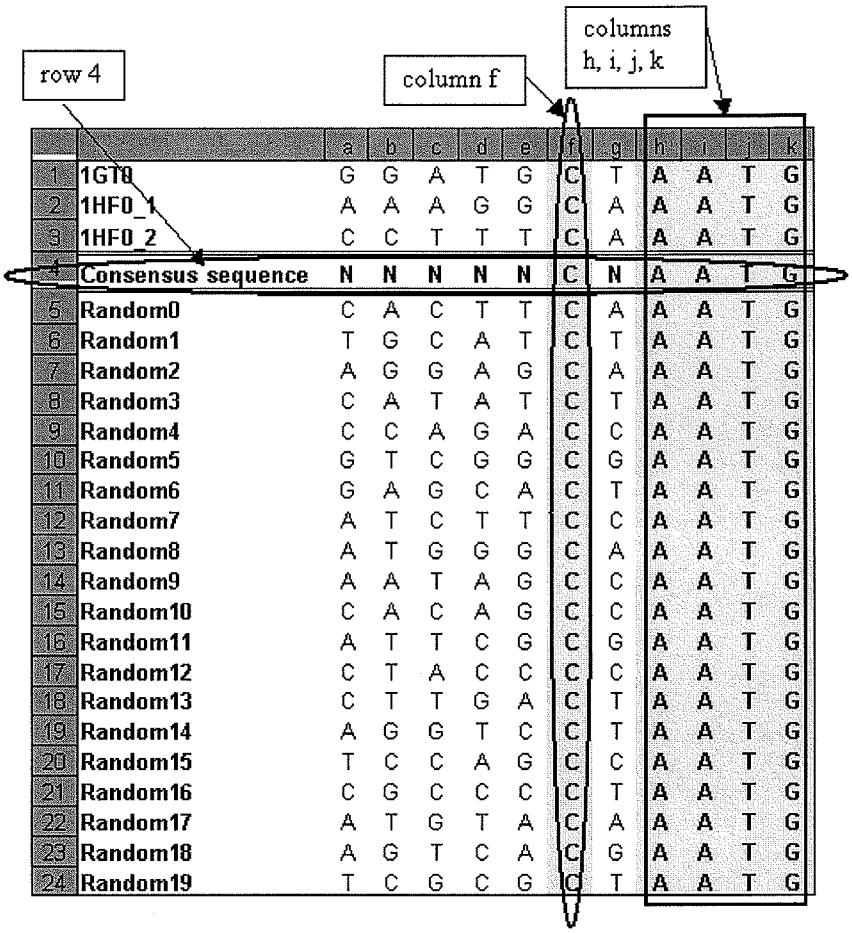


		a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p
1	1GTW	A	A	T	G	T	G	G	C	G	C	A	A	T	C	C	T
2	1GU4	A	A	T	A	T	T	G	C	G	C	A	A	T	C	C	T
3	1GU5	A	T	G	A	T	T	G	G	C	C	A	A	C	A	C	A
4	1H88	A	G	G	A	T	T	G	C	G	C	C	A	C	A	T	C
5	Consensus sequence	A	N	N	N	T	N	G	N	N	C	N	A	N	N	N	N
6	Random0	A	A	C	A	T	A	G	T	C	C	T	A	C	G	G	T
7	Random1	A	G	T	T	T	T	G	C	C	C	A	A	C	G	T	C
8	Random2	A	T	G	C	T	G	G	G	T	C	A	A	C	C	A	A
9	Random3	A	A	A	A	T	T	G	T	T	C	T	A	G	G	T	T
10	Random4	A	C	A	G	T	T	G	C	C	C	C	A	T	A	T	G
11	Random5	A	T	C	T	T	T	G	T	A	C	A	A	T	C	C	G
12	Random6	A	C	A	A	T	T	G	A	A	C	A	A	T	G	A	T
13	Random7	A	A	C	G	T	T	G	G	A	C	T	A	A	G	T	G
14	Random8	A	G	C	C	T	G	G	C	G	C	C	A	T	A	T	T
15	Random9	A	A	A	T	T	A	G	T	A	C	C	A	G	A	T	G
16	Random10	A	G	A	G	T	C	G	C	T	C	T	A	C	A	G	G
17	Random11	A	T	G	A	T	A	G	G	T	C	G	A	A	T	T	A
18	Random12	A	T	G	A	T	G	G	T	C	C	A	A	T	T	G	C
19	Random13	A	G	G	C	T	A	G	C	C	C	C	A	G	T	T	T
20	Random14	A	T	A	T	T	G	G	C	G	C	T	A	C	A	A	C
21	Random15	A	T	A	G	T	A	G	G	T	C	G	A	C	A	C	G
22	Random16	A	G	G	T	T	T	G	C	C	C	C	A	T	C	T	C
23	Random17	A	A	G	T	T	G	G	G	G	C	T	A	T	C	A	G
24	Random18	A	A	G	C	T	A	G	G	C	C	G	A	T	C	G	G
25	Random19	A	A	G	G	T	T	G	G	C	C	A	A	G	G	T	T

[0068] Table 4 displays the Consensus Sequence for "CAAT/Enhancer Binding Protein Beta" in row 5 and the nineteen (19) random sequences Random[0 – 19] that were generated by using conserved nucleotide bases in columns a, e, g, j and l, and either of the 4 possible bases in the rest of the columns, wherein each was randomly generated with equal probabilities.

[0069] Table 4. Consensus Sequence for "CAAT/Enhancer Binding Protein Beta"

[0070]



		a	b	c	d	e	f	g	h	i	j	k
1	1GT0	G	G	A	T	G	C	T	A	A	T	G
2	1HF0_1	A	A	A	G	G	C	A	A	A	T	G
3	1HF0_2	C	C	T	T	T	C	A	A	A	T	G
4	**Consensus sequence**	**N**	**N**	**N**	**N**	**N**	C	**N**	A	A	T	G
5	**Random0**	C	A	C	T	T	C	A	A	A	T	G
6	**Random1**	T	G	C	A	T	C	T	A	A	T	G
7	**Random2**	A	G	G	A	G	C	A	A	A	T	G
8	**Random3**	C	A	T	A	T	C	T	A	A	T	G
9	**Random4**	C	C	A	G	A	C	C	A	A	T	G
10	**Random5**	G	T	C	G	G	C	G	A	A	T	G
11	**Random6**	G	A	G	C	A	C	T	A	A	T	G
12	**Random7**	A	T	C	T	T	C	C	A	A	T	G
13	**Random8**	A	T	G	G	G	C	A	A	A	T	G
14	**Random9**	A	A	T	A	G	C	C	A	A	T	G
15	**Random10**	C	A	C	A	G	C	C	A	A	T	G
16	**Random11**	A	T	T	C	G	C	G	A	A	T	G
17	**Random12**	C	T	A	C	C	C	C	A	A	T	G
18	**Random13**	C	T	T	G	A	C	T	A	A	T	G
19	**Random14**	A	G	G	T	C	C	T	A	A	T	G
20	**Random15**	T	C	C	A	G	C	C	A	A	T	G
21	**Random16**	C	G	C	C	C	C	T	A	A	T	G
22	**Random17**	A	T	G	T	A	C	A	A	A	T	G
23	**Random18**	A	G	T	C	A	C	G	A	A	T	G
24	**Random19**	T	C	G	C	G	C	T	A	A	T	G

[0071] Then the Jack knife test was used to test the effectiveness of using the surface charge pattern representation to predict/identify DNA binding sites. For the simplicity of comparison and verification, three consensus matrices for "Octomaer-Binding Transcription Factor 1" were generated by omitting one actual DNA binding sequence, wherein the omitted sequence would be tested among the random samples, as shown in FIG 13. Similarly, two consensus matrices for "CAAT/Enhancer Binding Protein Beta" were generated and tested as shown in FIG 14.

[0072] FIG 13 shows the result of the test of "Octamer-Binding Transcription Factor 1" DNA binding sequences. As mentioned above, in each of the tests, one of the known DNA binding sequences (**) was left out in the generation of consensus matrix. The rest of the known DNA binding sequences (*) were used to generate the consensus matrix. The resultant consensus matrices were used to score all test sequences. The tables are sorted in descending order of the Score columns. The higher the score, the more likely the sequence is to bind to the protein. For TEST 1, using existing representation of A, C, T and G, 1H88 was ranked 6th, but indistinguishable from the other 5 randomly generated

test sequences with the same scores (Rand[1, 2, 6, 12, 17]). However, in the same test, using the surface charge representation of the present invention, 1H88 was ranked 4th and had similar scores with only one randomly generated test sequence Rand5. Thus the use of the proposed surface charge representation performed better than existing representations. Searches using the proposed surface charge representation were consistently better than existing nucleotide based representations in all the tests, as shown in FIG 13.

[0073] FIG 14 shows the result of the test of "CAAT/Enhancer Binding Protein Beta" DNA binding sequences. Similar to the results of "Octamer-Binding Transcription Factor 1" as described above, searches using the proposed surface charge representation were consistently better than existing representations in all the tests.

[0074] It is to be appreciated that while the surface charge patterns for single nucleotides and di-nucleotides are illustratively derived from the surface charges of the major grooves of the nucleotides, the principles of the representing a DNA sequence by using its surface charge patterns are not so limited. The surface charge patterns that are applicable in the present invention include, but are not limited to, the ones of the minor grooves of the nucleotides and the ones of any other views of the nucleotides.

[0075] While the present invention has been described with reference to particular embodiments, it will be understood that the embodiments are illustrative and that the invention scope is not so limited. Alternative embodiments of the present invention will become apparent to those having ordinary skill in the art to which the present invention pertains. Such alternate embodiments are considered to be encompassed within the spirit and scope of the present invention. Accordingly, the scope of the present invention is described by the appended claims and is supported by the foregoing description.

CLAIMS

What is claimed is:

1. A method of DNA sequence representation by using surface charge patterns, comprising:

obtaining a DNA sequence represented by linear nucleotide bases, N_1 N_2 N_3 N_4 ...N_{x-1} N_x, x = the total number of the nucleotide bases of the DNA sequence; wherein N is any nucleotide base selected from the group of A, T, G, and C;

adding 'di-nucleotides' into every inter-nucleotide space so that the DNA sequence may be represented by N_1 N_1-N_2 N_2 N_2-N_3 N_3 N_3-N_4 N_4 ...N_{x-1} N_{x-1}-N_x N_x, wherein N_{x-1}-N_x is any 'di-nucleotides' selected from the group of AA, AC, AG, AT, CA, CC, CG, CT, GA, GC, GG, GT, TA, TC, TG and TT; and

substituting surface charge patterns of each nucleotide base and 'di-nucleotides' with the DNA sequence N_1 N_1-N_2 N_2 N_2-N_3 N_3 N_3-N_4 N_4 ...N_{x-1} N_{x-1}-N_x N_x; thereby the DNA sequence is represented by a sequence of surface charge patterns.

2. The DNA sequence representation method of claim 1, wherein the surface charge patterns of each single nucleotide base are derived from the surface charges of the major groove of the each single nucleotide base within its corresponding pair; and wherein the surface charge patterns of each 'di-nucleotides' are derived from the surface charges of the major groove of the each 'di-nucleotides'.

3. The DNA sequence representation method of claim 2, wherein the surface charge patterns may be derived by applying a plurality of windows to the major grooves of the each single nucleotide base or the each 'di-nucleotides' and assigning with each window a value for the charge types and/or intensities of the surface charges of the each single nucleotide base or the each 'di-nucleotides', thereby the surface charge patterns will be represented by a string of symbols representing the values of the charge types and/or intensities of the surface charges of the major groove of the each single nucleotide base or the each 'di-nucleotides'.

4. The DNA sequence representation method of claim 3, wherein the symbols may be positive integers and/or negative integers and/or decimals.

5. A DNA sequence represented by a sequence of surface charge patterns including surface charge patterns of each single nucleotide base and 'di-nucleotides'; wherein the each single nucleotide base is selected from the group of A, T, G, and C; wherein the 'di-nucleotides' is selected from the group of AA, AC, AG, AT, CA, CC, CG, CT, GA, GC, GG, GT, TA, TC, TG and TT; and wherein the 'di-nucleotides' are inserted into the inter-space between two adjacent nucleotide base and consist of the two nucleotide bases abutting the insertion point.

6. The DNA sequence of claim 5, wherein the surface charge patterns of each single nucleotide base are derived from the surface charges of the major groove of the each single nucleotide base within its corresponding pair; and wherein the surface charge patterns of each 'di-nucleotides' are derived from the surface charges of the major groove of the each 'di-nucleotides'.

7. The DNA sequence of claim 6, wherein the surface charge patterns may be derived by applying a plurality of windows to the major grooves of the each single nucleotide base or the each 'di-nucleotides' and assigning with each window a value for the charge types and/or intensities of the surface charges of the each single nucleotide base or the each 'di-nucleotides', thereby the surface charge patterns will be represented by a string of symbols representing the values of the charge types and/or intensities of the surface charges of the major groove of the each single nucleotide base or the each 'di-nucleotides'.

8. The DNA sequence of claim 7, wherein the symbols may be positive integers and/or negative integers and/or decimals.

9. A method for identifying a DNA binding site, comprising:

aligning at least one known DNA binding site sequence, wherein the at least one known DNA binding site sequence is represented by a sequence of surface charge patterns

including surface charge patterns of each single nucleotide base and 'di-nucleotides'; wherein the each single nucleotide base is selected from the group of A, T, G, and C; wherein the 'di-nucleotides' is selected from the group of AA, AC, AG, AT, CA, CC, CG, CT, GA, GC, GG, GT, TA, TC, TG and TT; and wherein the 'di-nucleotides' are inserted into the inter-space between two adjacent nucleotide base and consist of the two nucleotide bases abutting the insertion point;

deducing a consensus representation of the aligned at least one DNA binding site sequence; and

comparing an unknown DNA sequence with the consensus representation; thereby the closer the unknown DNA sequence resembles the consensus representation, the more likely the unknown DNA sequence is a potential DNA binding site.

10. The method of claim 9, wherein the surface charge patterns of each single nucleotide base are derived from the surface charges of the major groove of the each single nucleotide base within its corresponding pair; and wherein the surface charge patterns of each 'di-nucleotides' are derived from the surface charges of the major groove of the each 'di-nucleotides'.

11. The method of claim 10, wherein the surface charge patterns may be derived by applying a plurality of windows to the major grooves of the each single nucleotide base or the each 'di-nucleotides' and assigning with each window a value for the charge types and/or intensities of the surface charges of the each single nucleotide base or the each 'di-nucleotides', thereby the surface charge patterns will be represented by a string of symbols representing the values of the charge types and/or intensities of the surface charges of the major groove of the each single nucleotide base or the each 'di-nucleotides'.

12. The method of claim 11, wherein the symbols may be positive integers and/or negative integers and/or decimals.

13. The method of claim 9, wherein the consensus representation may be any one selected from the group consisting of consensus sequence, consensus matrices, consensus

weighted matrices, weighted matrices of charges, pattern recognition method, and method of artificial intelligence.

14. The method of claim 9, wherein the consensus representation is a consensus weighted matrices of charges so that the closer the summation score of the unknown DNA sequence is to the maximum score of the consensus representation, the more likely the unknown DNA sequence is a potential DNA binding site.