5632–5640 *Nucleic Acids Research, 2009, Vol. 37, No. 17* *Published online 3 August 2009*
doi:10.1093/nar/gkp628

Local combinational variables: an approach used in DNA-binding helix-turn-helix motif prediction with sequence information

Wenwei Xiong, Tonghua Li*, Kai Chen and Kailin Tang

Department of Chemistry, Tongji University, Shanghai, 200092, China

Received April 28, 2009; Revised and Accepted July 14, 2009

ABSTRACT

Sequence-based approach for motif prediction is of great interest and remains a challenge. In this work, we develop a local combinational variable approach for sequence-based helix-turn-helix (HTH) motif prediction. First we choose a sequence data set for 88 proteins of 22 amino acids in length to launch an optimized traversal for extracting local combinational segments (LCS) from the data set. Then after LCS refinement, local combinational variables (LCV) are generated to construct prediction models for HTH motifs. Prediction ability of LCV sets at different thresholds is calculated to settle a moderate threshold. The large data set we used comprises 13 HTH families, with 17 455 sequences in total. Our approach predicts HTH motifs more precisely using only primary protein sequence information, with 93.29% accuracy, 93.93% sensitivity and 92.66% specificity. Prediction results of newly reported HTH-containing proteins compared with other prediction web service presents a good prediction model derived from the LCV approach. Comparisons with profile-HMM models from the Pfam protein families database show that the LCV approach maintains a good balance while dealing with HTH-containing proteins and non-HTH proteins at the same time. The LCV approach is to some extent a complementary to the profile-HMM models for its better identification of false-positive data. Furthermore, genome-wide predictions detect new HTH proteins in both *Homo sapiens* and *Escherichia coli* organisms, which enlarge applications of the LCV approach. Software for mining LCVs from sequence data set can be obtained from anonymous ftp site ftp://cheminfo.tongji.edu.cn/LCV/freely.

INTRODUCTION

Since the discovery of the DNA double helix in 1953 and the 'central dogma' of molecular biology (1), which has been questioned and subsequently revised, research and debate on the flow of genetic information have been continuous (2). The mechanisms for encoding, decoding and transmitting genetic information have been the focus of much attention. DNA-binding proteins play a vital role in the delivery of this information DNA-binding proteins (3) include transcription factors that modulate the transcription process, nucleases that cleave DNA molecules and histones that are involved in DNA packaging in the cell nucleus (4). DNA-binding proteins comprise DNA-binding domains such as the helix-turn-helix (HTH), the zinc finger and the leucine zipper, among others (5). Since the determination of the crystal structures of C1 and Cro repressor proteins from the lambda bacteriophage, the DNA-binding HTH structural motif has become one of the most important studied examples of the interaction between proteins and DNA (6).

The HTH structural motif is composed of two helices joined by a short strand of amino acids and is found in many proteins that regulate gene expression. In most cases, such as in the Cro repressor, the second helix contributes most to DNA recognition, and hence it is often called the recognition helix. It binds to the major groove of DNA through a series of hydrogen bonds and various van der Waals' interactions with exposed bases (7).

Prediction methods for determining whether proteins contain the HTH motif represent a hot topic, and a number of prediction methods have been proposed in recent years. Structure-based methods include HTHquery, a web-based service based on a similarity with a set of structural templates, the accessibility of a putative structural motif and a positive electrostatic potential in the neighborhood of the putative motif (8); use of the electrostatic potential to select generic DNA-binding residue patches (9) and a statistical model based on geometric measures of the

*To whom correspondence should be addressed. Tel: +86 21 65983987; Email: lith@tongji.edu.cn

motif with a decision tree model (10). Apart from several consensus-based and profile-based approaches dating back to the 1990s or earlier and a number of evolutionary studies, few methods using only the sequence information have been published. For instance, a fully connected two-layered neural network for a series of structural and sequence features was proposed for the prediction of DNA-binding proteins and residues (11). In another good work using both structural knowledge and sequence information, two different libraries of hidden Markov models were built, results showing that information carried by motif sequences and motif structures are to some extent complementary (12).

Despite the growing number of proteins for which the 3D structure has been determined, only primary sequence information is available for many proteins. Thus, it is essential to develop an effective way to predict HTH structure motifs using only primary sequence information.

Variable extraction and present of samples (i.e. protein primary sequence) are of critical importance for biostatistical reasoning and modeling. These serve as the information bridge from samples to models that may reveal potential biological meanings. Many studies attach importance to variable extraction and calculate frequency statistics for samples that are directly used or transformed to other formats for constituting feature variables. For better prediction, the variables extracted for depicting the data set should differ greatly between the positive set and the negative set (13). A study that used the frequency difference between true sites and false sites to predict splice sites is a case in point (14).

When extracting variables from samples of different length, strategies must be used to process the raw data to project the same multi-dimensional feature space. Some traditional methods used include approaches that consider amino acid frequency encoding, the composition, transition and distribution of amino acids (15), and the physical and chemical properties of amino acids (16). These methods show good performance for some specific areas, but do not yield satisfactory results for all cases.

HTH motif prediction remains a challenge, because the motifs are much shorter than the sequences containing them, unique local information from other parts of the sequence cannot be effectively utilized. Furthermore, the position of the HTH motif varies considerably from protein to protein. It is hard to focus on motif domains when considering the whole protein sequence. Residues contributing to the same frequency-based variable may have many possible permutations and combinations at the primary sequence level. Thus, traditional methods using frequency statistics show weakness in extracting local structural information for proteins and generating effective variables to construct a good model.

Avoiding overall encoding of protein sequences, methods based on profile hidden Markov models (profile HMMs) are state of the art. In this method, protein domains are divided into families (or clans), which are related sequences defined by similarity of sequence, structure or profile-HMM. An alignment of a set of representative sequences (usually named SEEDs) in a clan is used to construct the HMMs. Pfam (17) is such a good protein families database including comprehensive collection of protein domains and families, represented as multiple sequence alignments and as profile hidden Markov models (22). Proteins with unknown structure information can be submitted to Pfam server to detect whether a HTH motif is contained and which clans it belongs to. The HTH motif detection ability of Pfam is powerful due to its detailed classification; however it is inferior in discriminating non-HTH proteins that have higher sequence similarity to SEEDs from HTH clans (18).

We developed a new approach based on a quintessential HTH motif set to generate local combinational variables (LCVs). This method showed good performance for large-scale HTH motif prediction using only primary sequence information. First, a quintessential set containing 88 peptides of 22 residues was used to generate local combinational segments (LCSs). Then a traversal algorithm was carried out to exhaust all possible residue combination patterns. This step is time-consuming because of the number of possibilities for residue positions and counts. To minimize this process, some rules were introduced to enhance the algorithm efficiency. After a mining step, LCSs were used in a sliding window matching process to generate LCVs, which were consequently taken as the feature variables of the motif prediction model. Using the SMART database of 13 HTH-motif-containing protein families, we verified the effectiveness of the LCV approach for HTH prediction, with overall results of 93.29% accuracy, 93.93% sensitivity and 92.66% specificity. Better prediction results of newly reported HTH-containing proteins compared with other prediction web service (i.e. GYM2.0) validate generalization ability of the prediction model. The LCV approach was also compared to the profile-HMM method of Pfam, and the former maintained a good balance between HTH-containing proteins and non-HTH proteins while the latter showed powerful detection ability of known HTH-containing proteins. Better discrimination of false positive data proved a balance prediction model derived from LCV approach, which could be a complementary to profile-HMMs. Furthermore, genome-wide predictions detect new HTH proteins in both *Homo sapiens* and *Escherichia coli* organisms, which enlarge applications of the LCV approach.

MATERIALS AND METHODS

The LCV approach is an encoding method for protein sequence which extracts LCS from an aligned protein dataset named quintessential set at a moderate threshold, and generates LCV by matching refined LCS set to target proteins. These local combinational variables can be used in machine learning algorithms to set up models and make predictions.

Data sets

Three data sets, QuintessentialSet-88, DS_HTH_ALL and DS_NONHTH_ALL (see below for definition), were used to construct prediction models and to compare the new LCV approach with some traditional methods.

Table 1. HTH families

HTH family	Sequence	Description from SMART
HTH_ARAC	5426	Arabinose operon control protein
HTH_ARSR	2111	Arsenical resistance operon repressor
HTH_ASNC	1571	An autogenously regulated activator of asparagine synthetase
HTH_CRP	810	cAMP regulatory protein
HTH_DEOR	1099	Deoxyribose operon repressor
HTH_DTXR	252	Diphteria tox regulatory element
HTH_GNTR	4402	Gluconate operon transcriptional repressor
HTH_ICLR	1058	Isocitrate lyase regulation
HTH_LACI	2137	Lactose operon repressor
HTH_LUXR	4068	Lux regulon
HTH_MARR	2777	Multiple antibiotic resistance protein
HTH_MERR	2046	Mercury resistance
HTH_XRE	6002	XRE-family like proteins

QuintessentialSet-88 comprises 88 protein sequences (19) of 22 amino acids in length. The sequences are aligned and are non-redundant. Choosing a quintessential set is vital for the remaining steps. Initially, research on HTH motif prediction was based on scores between a master set and a target protein sequence. In this approach it was assumed that similarity to known HTH motif sequences indicated whether a protein contained HTH motifs. Thus, master sets have been supplemented with increasing experiment data (7,20). The quintessential set we used was the master set of Dr Giri Narasimhan in his GYM2.0 program (19). The proteins are chosen from different species and have different functions.

DS_HTH_ALL data set contains 17 455 sequences (up to November 1, 2006), each containing at least one HTH motif. By browsing the SMART database (21), we get 13 HTH families (Table 1). Each item has its accession number (ACC), by which raw protein sequence can be obtained from other protein databases; in the present study we used the ExPASy proteomics server (22).

DS_NONHTH_ALL contains the same number (i.e. 17 455) of proteins as DS_HTH_ALL, which are all from ExPASy database without any reported HTH motifs. In the modeling process, it is taken as negative data set. In modeling of each HTH motif family, sequences were randomly chosen with the same number as corresponding positive data set.

Traditional encoding methods

Some traditional methods were tested for construction of prediction models for HTH motifs:

Single-residue frequency method (23): the content of 20 amino acids in protein sequences is calculated;

Double-residue frequency method (14): the content of all possible double-residue combinations (20×20) is calculated and

Composition–transition–distribution method (CTD) (15): the composition is calculated as a percentage of three constituents/groupings (e.g. polar, neutral and hydrophobic residues for the feature of hydrophobicity). The transition frequencies (polar to neutral, neutral to hydrophobic, etc.) are then calculated. Finally, the distribution pattern of a particular property (the position of the first amino acid with a given property and the sections in which 25, 50, 75 and 100% of the amino acids with that property are contained) is determined.

LCV concept

The definitions and concept of the LCV approach are as follows.

The quintessential set contains several protein sequences that are non-redundant, of the same length and from different protein families to ensure a representative distribution of HTH motifs. A residue and its position in the sequence, e.g. A3, are termed the basic unit of the LCS. An LCS consists of several such units and is depicted as the LCS units in brackets, e.g. {A3, E5, F8}. An LCS containing only one unit is called an LCS seed. Some LCSs may be a part of others, e.g. {A3, D5} is part of {D2, A3, D5, F7}, where the former is a sub-LCS of the latter.

If an LCS appears in a sequence (e.g. {A3, D5, F6} in '_ _A_DF_', where _ denotes any amino acid), the LCS matches or supports the sequence. When sequences in the quintessential set are aligned, we may obtain many LCSs for each sequence, with some LCSs simultaneously supporting many different sequences. These simultaneous LCSs show the same amino acid composition in different sequences, and we can calculate the number of times they appear, named support number.

To constrain the number of LCSs, we set a threshold to filter out LCSs with a support number lower than the threshold. Then we use criteria to refine and optimize the LCS set. The LCS set for a given quintessential set and threshold is thus defined. The LCS matches the target data set in a different way compared to the quintessential set. Whereas an LCS matches a sequence in the quintessential set with absolute positions (e.g. {A3, E5} only matches sequences like '_ _A_E_'), it matches the target data set with relative positions (e.g. {A3, E5} matches both '_ _A_E_' and '_A_E_ _').

After LCS refinement and optimization, the match counts to the target data set (i.e. protein sequences that contain HTH motifs or not) for each LCS are calculated as the LCV variables. Because of the difference in protein length between the quintessential set and the training set, a successful match occurs when a certain LCS matches the target sequence in relative position. An LCS-length window slides along the target sequence to detect successful matches. The match count for the LCS set for each sequence in the target data set is used to generate the LCV. Thus, the LCV indicates the sequence accordance of the target data set with the LCS set, and the LCS incidence.

Finally, support vector machine (SVM) tools are used to construct a prediction model from the LCVs.

LCS mining algorithm rules

To utilize the information of the quintessential set effectively, some highly representational segments should be extracted from sequences. This is a time-consuming step because of the variation in residue combination

positions and counts. To increase the algorithm efficiency, the following rules are introduced. These rules used in the mining algorithm greatly improve the time complexity. They are relevant for a given quintessential set and a certain threshold value (N is the length of a given sequence).

(1) If and only if the support number for an LCS in a quintessential set is not less than the given threshold, the LCS is valid and can be retained in iteration steps.
(2) In a given quintessential set, the present number of an N-length LCS is surely more than all its sub-LCS.
(3) The sub-LCS of a given valid LCS is also a valid LCS.
(4) An N-length LCS is the result of merging of two $N-1$-length sub-LCSs that have $N-2$ similar residues when $N>2$.
(5) This rule is also known as the LCS merging rule. When $N>1$, two N-length LCSs can be merged if and only if they have $N-1$ LCS units the same and one different LCS unit at different positions. When $N=1$ (LCS seed), two N-length LCSs can be merged into one $N+1$-length LCS if and only if the seeds are at different positions. It depends on the match counts whether the new LCS is valid.

Algorithm implementation

The algorithm is implemented as follows.

(1) Quintessential set choice and initialization

A set of aligned protein sequences of the same length is needed to generate the LCS set, which should be non-redundant and to some extent typical, otherwise, the LCVs generated may be not suitable. In the present study, QuintessentialSet-88 meets these requirements.

The number of LCSs generated depends on the threshold chosen in the algorithm. A moderate value is essential to provide the prediction model with enough information and not too big variable dimension.

(2) LCS seed searching

The search for LCS seeds process is a time-consuming step. All sequences of the same length form a residue matrix. Each type of residue in each column is counted as the present number. If this is not less than the target threshold, the residue is eligible as an LCS seed.

(3) Iteration process.

Based on rules 1 and 5 and using seeds obtained from the last step, the iteration process can be carried out until the longest LCSs are generated (i.e. of the same length as the protein sequence), or no more LCSs are generated in a certain iteration. To minimize the algorithm complexity, each LCS position should be marked.

Cross-validation of the HTH motif prediction model

For each HTH motif family (denoted HTH_XXX) a 5-fold cross-validation was performed to test the HTH motif prediction model. The positive data set consisted of the sequence in the central 80% distribution of all sequences in length. The same number of sequences randomly chosen from DS_NONHTH_ALL constituted the negative data set.

SVM tools are widely used in machine learning for supervised classification and regression. We use LibSVM (24), which is an integrated software for support vector classification, for all experiments. A radial basis function was chosen as the kernel function to construct prediction models.

RESULTS

Results for traditional methods

The prediction results for the best traditional methods are shown in Table 2 for various model parameters.

Prediction ability for different thresholds

A moderate threshold is important for generating a suitable number of LCSs and LCVs. A threshold that is too low will result in too many LCSs and a long computation time. On the other hand, a threshold that is too high will result in too few LCSs and LCVs, and there will not be enough information to construct the prediction model. The prediction ability for different thresholds is listed in Table 3 and the performance of the prediction model for a moderate threshold of 7 is shown in Supplementary Data Table 1.

LCV results for different thresholds

We tested different thresholds and obtained different numbers of LCVs. The cross-validation results are shown in Supplementary Data Tables 1–4. From these results it can be concluded that in a certain range, the lower the threshold and the greater the number of LCVs, the better is the result. Moreover, when threshold

Table 2. Results for traditional methods

Method	Variable count	Ac (%)	Sn (%)	Sp (%)	CC
Single-residue	20	74.50	70.59	78.41	0.49
Double-residue	400	80.65	76.91	84.39	0.61
CTD	104	76.72	73.92	79.51	0.54

Ac, Sn, Sp and CC denote the accuracy, sensitivity, specificity and correlation coefficient, respectively.

Table 3. Prediction ability for different thresholds

Threshold	Variable count	Ac (%)	Sn (%)	Sp (%)	CC
7	567	93.84	94.85	92.82	0.88
8	354	93.29	93.93	92.66	0.87
13	78	88.20	89.59	86.80	0.77
17	27	83.21	84.80	79.63	0.65

Ac, Sn, Sp and CC denote the accuracy, sensitivity, specificity and correlation coefficient, respectively.

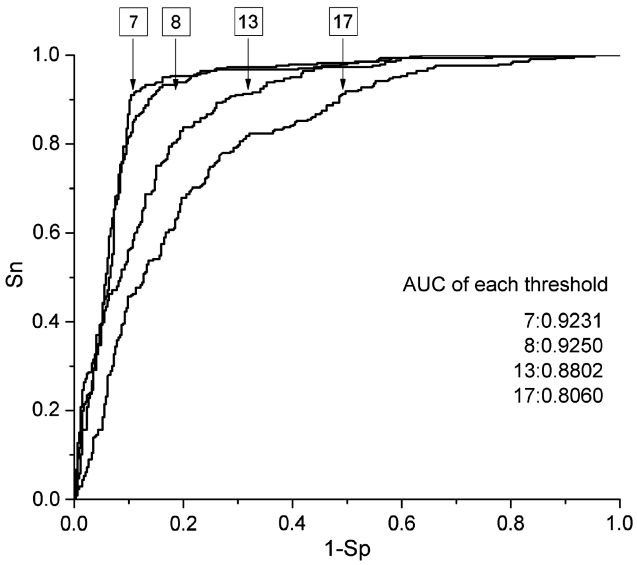

Figure 1. ROC curve for different thresholds (threshold of 7, 8, 13 and 17 from top to bottom).

varies from 8 to 7, the number of LCVs changes a lot (from 354 to 567), but there is no significant improvement in prediction ability, and the accuracy only increases from 93.29% to 93.84%. In comparison, when the threshold varies from 7 to 13, the number of LCVs decreases from 354 to 78, and the prediction ability also significantly decreases with the accuracy decreasing from 93.29% to 88.20% (Figure 1). Thus, a threshold of 8 is a moderate value.

Comparisons with GYM2.0

To verify the generalization capability of the 13 LCV-based prediction models, we implemented experiments on newly reported HTH-motif-containing sequences from the SMART database.

Sequences containing HTH motifs of 2 HTH families, i.e. HTH_ASNC and HTH_DEOR (see Table 1) in the SMART database up to now, are all tested except for those taken in former prediction models. The two new prediction models are constructed with all sequences used in former prediction models of each HTH family as positive set, and equal amount of sequences randomly chosen from DS_NONHTH_ALL as negative set. Some results are pretty good as the cross validation results, e.g. 95.05% accuracy in HTH_ASNC of 1274 sequences, while others are not so good as cross-validation results, e.g. 84.78% accuracy in HTH_DEOR of 736 new sequences. It is possible that sequences in HTH_DEOR family are not enough and not so typical that leads to the inferior generalization capability of its prediction model. Better results can be expected with sufficient HTH-motif-containing sequences. Compared to cross-validation, anyhow, these results finally confirm that the LCV approach is applicable in HTH motif prediction.

Compared with Dr Giri Narasimhan's work in his GYM2.0 web service (http://www.cs.fiu.edu/~giri/bioinf/GYM/welcome.html), results of GYM2.0 prediction are 66.88% accuracy for HTH_ASNC and 81.79% accuracy for HTH_DEOR (Table 4). The lower prediction accuracy may due to that HTH motifs are not always exactly 22 amino acids in length and fixed-length sliding window for matching motif sequences still has some limitations. Meanwhile, the LCV approach does not extract information from only fixed-length-part of a given sequence but involves all the 'contribution factors' for HTH motifs. In other words, more flexible utilization of sequence ensures more intensively information mining.

Table 4. Prediction results compared with GYM2.0

HTH family	Count	GYM detected	Ac (%)	LCV detected	Ac (%)
HTH_ARAC	4463	3492	78.24	3618	81.07
HTH_ARSR	1597	756	47.34	1240	77.65
HTH_ASNC	1274	852	66.87	1211	95.05
HTH_CRP	561	544	96.97	522	93.05
HTH_DEOR	736	602	81.79	624	84.78
HTH_DTXR	195	126	64.61	156	80.00
HTH_GNTR	3454	2701	78.20	3154	91.31
HTH_ICLR	760	580	76.32	706	92.89
HTH_LACI	1485	1454	97.91	1455	97.98
HTH_LUXR	2909	2331	80.13	2314	73.36
HTH_MARR	2152	1156	53.72	1934	89.87
HTH_MERR	1516	1103	72.76	1347	88.85
HTH_XRE	4454	3828	85.95	4164	93.49
Total	25 556	19 525	76.40	22 445	87.83

Table 5. Prediction results of LCV and Pfam

Method	Ac (%)	Sn (%)	Sp (%)	CC
LCVs (1)	95.40	92.60	96.45	0.88
HMM (1)	87.05	100.00	81.87	0.75
LCVs (2)	93.12	90.91	93.83	0.82
HMM (2)	69.90	97.15	61.10	0.50

LCVs (1) denotes LCVs derived from the QuintessentialSet-88 set; HMM (1) denotes the HMM model of Crp clan in Pfam database; LCVs (2) denotes LCVs derived from SEEDs in Crp clan and HMM (2) denotes HMM model constructed by sequences in the QuintessentialSet-88 set. In both HMMs, models are calibrated to increase the sensitivity of search, and *E*-values are empirical estimates.

Comparisons with Pfam

By searching profiled HMMs, Pfam uses sequence and domains score to determine whether a sequence belongs to the full alignment of a particular Pfam entry. The attribute of domains and motifs depends on sequence similarity, thus those non-motif sequences with certain degree of sequence similarity may result in incorrect discriminations.

To compare with the profile-HMM method, 561 newly reported proteins in HTH_Crp family are taken as positive test set. Proteins without HTH motifs are randomly chosen as negative test set. First, the LCV approach with threshold 8 and the Pfam clan Crp are both used on test set. The LCV approach shows a better balance between positive and negative sets (Table 5). Then the SEEDs from Crp clan are taken as the quintessential set to generate

different LCVs for HTH prediction. It is interesting that the new LCVs perform not so good as the QuintessentialSet-88 set, which may be due to the SEEDs form Crp clan share high sequence similarity and sequence of train set and test set may differ from the new LCVs to a larger extent compared with the original LCVs. Furthermore, when using the QuintessentialSet-88 to build profile-HMM model, the specificity of negative test set decreases significantly while the sensitivity of positive set becomes a little lower than its own SEEDs. It can be inferred that HTH SEEDs from different protein families combine characteristic information of each family, and then have more prediction accuracy on positive set but poor results on negative sets for its lower conservation in sequence similarity. While the profile-based HMM method has powerful ability in detecting known HTH-containing proteins, the LCV approach can take advantage of the combinational information effectively to build prediction models and yields balance results, and therefore can be regarded as a beneficial complementary to Pfam. In fact, the SEEDs in Pfam database can be regarded as combinations of the LCVs, which are arranged with certain sequence orders and fixed length. The LCV approach focus on those highly presented residue combinations from quintessential set, and ignores the sequence order of LCV location.

New predictions on genome-wide proteins

To make the LCV approach broadly applicable, predictions on a genomic scale are launched to find out new HTH-containing proteins. Proteins in two organisms (e.g. *H. sapiens* and *E. coli*) from the UnitProt Knowledgebase (UniProtKB) are investigated. UniProtKB (25) gives access to all the protein sequences which are available to the public with different evidence levels of protein existence. More than 99% of the protein sequences in UniProtKB are derived from the translation of the coding sequences (CDS) which have been submitted to the public nucleic acid database, the EMBL-Bank/GeneBank/DDBJ database.

There are totally 86 824 sequences for *H. sapiens* and 16 658 sequences for *E. coli*, of which 20 330 and 641 sequences are reviewed, respectively. First, in each organism, proteins recognized by Pfam database to have HTH motifs are taken as positive set of new prediction models (i.e. 497 HTH proteins for *H. sapiens* and 1172 HTH proteins for *E. coli*), while same amount of those non-HTH proteins are randomly chosen as negative set. Second, new prediction models are constructed by taking LCVs from QuintessentialSet-88 with a threshold of 8. In each organism, models are training for five times with different random selected negative set, and an average of 2% false positive rate and 99% sensitivity of self-prediction results show good model prediction ability and consistency. Then all proteins in both organisms are tested and those newly detected HTH-containing proteins are analyzed of their subcellular locations and gene ontology. 350 proteins against all reviewed proteins in *H. sapiens* organism are newly detected having HTH motifs and the same for 861 proteins against all proteins in *E. coli* organism (see Supplementary Data Table 6). Known subcellular locations and gene ontology distributions are listed in Table 6. As for DNA-binding and transcriptional regulating, HTH motifs are most probably appeared in nucleus, membrane and cytoplasm.

Table 6. Statistics on newly detected HTH proteins of subcellular location and gene ontology

Homo sapiens subcellular	Count	*H. sapiens* gene ontology terms	Count
Membrane	76	C:integral to membrane	88
Nucleus	70	C:nucleus	72
Secreted	51	F:protein binding	59
Cytoplasm	48	C:cytoplasm	41
Cell membrane	21	P:transcription	36
Endoplasmic reticulum membrane	16	P:regulation of transcription, DNA-dependent	35
Mitochondrion	10	C:extracellular region	33
Extracellular space	9	F:transcription factor activity	28
Cell junction	5	F:zinc ion binding	23
Total	387	Total	1477
Escherichia coli subcellular	Count	*E. coli* ontology terms	Count
Cytoplasm	10	F:DNA binding	43
Periplasm	7	P:regulation of transcription, DNA-dependent	37
Fimbrium	1	C:membrane	28
Secreted	1	P:signal transduction	20
		P:chemotaxis	19
		F:transcription factor activity	18
		F:transmembrane receptor activity	18
		P:pathogenesis	14
		P:transposition, DNA-mediated	13
		P:DNA recombination	13
Total	19	Total	592

There are 350 and 861 new HTH proteins detected in *H. sapiens* and *E. coli* organism, respectively. Total count under each item is the count of annotation entries that can be found in UnitProtKB database. Some reviewed proteins may have more than one annotation entries; in contrast, unreviewed ones may have no annotations. Only the top 10 annotation entry groups are listed.

DISCUSSION

LCV's advantages

The LCV approach takes into account non-consecutive residues rather than involves specific properties of protein entirely as the traditional methods. Due to the gaps among residues in LCVs, it is almost impossible to generate the same variables by using traditional frequency statistics methods, because of the rapid exponential growth of variable count while increasing the statistical residues and interval. On the other hand, the LCV approach encoding samples of unequal length to equal length variables can effectively make use of sequence information and avoid exhausting all possible combinations of residues.

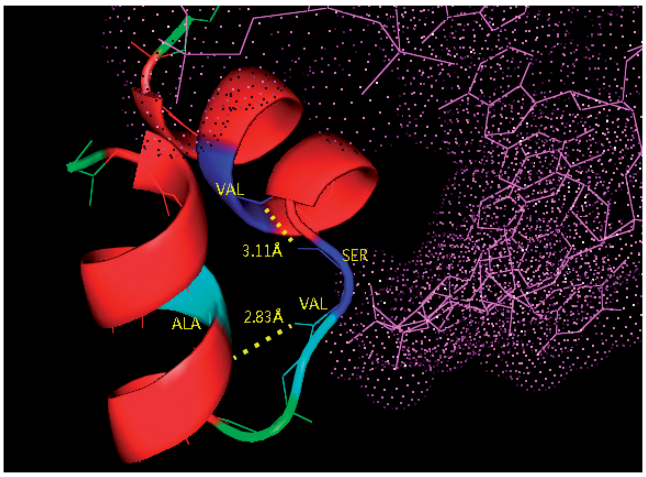

Figure 2. Long-range interactions between residues in LCV within the HTH motif. The HTH motif part (red for α-helix and green for turn) and its binding DNA (purple part) are shown in 3D structure of molecule (i.e. PDB Id:1BDI). Two potential interactions occur between residues of the {A7,V12} (cyan part) and {S13,V17} (blue part) LCVs, of which the distances are 2.83 and 3.11 Å, respectively.

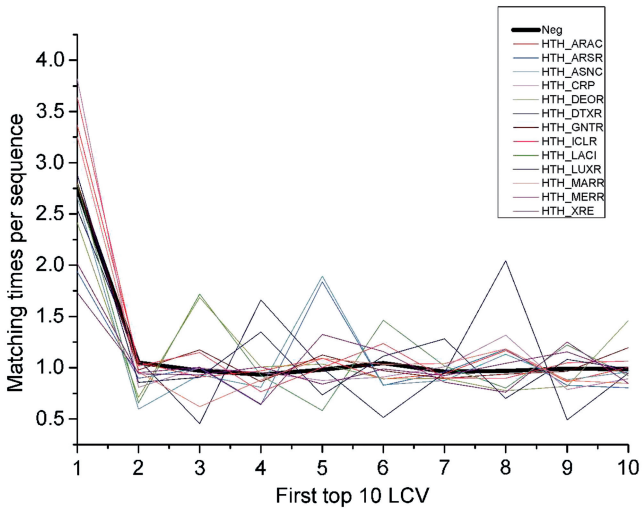

Figure 3. LCV count distribution for each HTH family. The number of matches per sequence is shown for the top 10 LCV numbers.

Potential long-range interaction between residues

In addition to LCV's advantages mentioned above, regarding the inconsistency between primary sequence and 3D structure, traditional methods based on residue statistics can hardly involve variables reflecting long-range interactions, while LCVs may include residues which potentially interact with each other, owing to the non-consecutive feature of LCV. From distance analysis, residues in specific LCVs drop in a possible interaction distance. These potential long-range interactions may be an important factor in formation of HTH motifs.

Proteins containing HTH motifs with known 3D structures in the Protein Data Bank database are investigated. LCVs in the motif region of proteins are possible combinations of residues which have potential long-range interactions. By calculating distances of main-chain atoms, some residues having potential long-range interactions are singled out. An example in a DNA-binding protein complex (i.e. PDB Id: 1BDI) is shown in Figure 2. Two residue groups highlighted by cyan and blue colors appear both in LCVs (i.e. {A7,V12} and {S13,V17}) and in HTH motifs, with distances of 2.83 and 3.11Å, respectively. It can be seen from the structure that the residues in both α-helix may interact with residues in the turn secondary structure according to the close distances.

LCV refinement and optimization

LCS seeds should be removed from the LCS set, since these are single residues at specific positions and thus cannot represent the combinational relationship among residues.

There may be some redundant LCSs that play the same role in generating LCVs. For example, two LCSs, {A1,E2,D6} and {A3,E4,D8}, are the same in the window sliding process. Only one should be retained for LCV generation.

When the number of matches for an LCS and its sub-LCS are equal, the shorter one should be removed. If the sub-LCS has more matches than the parent LCS, it should be retained. This measure avoids an excessive number of variables, and at the same time not neglects useful information.

LCV distribution in each HTH family

The LCV count distribution was calculated for each protein family, as shown in Figure 3. It can be concluded that some LCVs are more prominent in all the protein families, whereas others show a different distribution, and the top 10 LCVs are distributed more evenly in the negative set than for other HTH families. It may be the varying LCV distribution for each HTH family that contributes to the better prediction ability.

LCS coverage analysis

A good LCS set should represent all the protein sequence information in the quintessential set. The LCS coverage indicates the utilization ratio of the quintessential set. The ideal LCS coverage is 100%; a lower value indicates that information is missing and the prediction model may be weak. The LCS set used in this work reaches 100% coverage of the quintessential set.

From the perspective of specific locations of aligned-quintessential set, the location coverage reveals importance of 22 locations within HTH motifs. The 7th, 11th and 17th locations have higher percentage of residues. It should be noted that in the 7th location amino acid Ala (A) covers all the cases and Gly (G) almost the same in the 11th location. What's more interesting are, Ala has the highest helix propensity while Gly has the lowest helix-forming propensity and tends to disrupt helices because its high conformational flexibility makes it entropically expensive to adopt the relatively constrained α-helical structure (26).

LCV extensibility and applicability

There are many structural domains or motifs in different length distributed widely in various proteins which need to be predicted with primary sequence information. The LCV approach shows effectiveness on the HTH motif with an average length of 22 residues. Some shorter motifs (e.g. beta turn with four residues) may not take advantage of the LCV approach for its less number of variable combinations. Since most known domains have an average length of more than 20 residues (25), domain size does not confine extensible use of LCV. Cases are more complicated with bigger domains, which do not have a single conserved motif but several discontinuous motifs positioned with variable spacings, with a different degree of conservation and sometimes with different order of motifs in linear sequence. A case in point is enzyme prediction and its functional classification. Efforts have been made to identify whether a newly found protein sequence is an enzyme or not, and if it is, to determine which main class does it belong to (27). Beyond the previous works, which take into account pair-wise sequence similarity (28), physical and chemical features (29), and functional domain composition and pseudo amino acid composition (27), a good work done by Kunik *et al.* (30), introduced the Specific Peptides to perform function annotation of enzyme and achieved better results. Both LCV and Specific Peptide are deterministic combinations of amino acids. While Specific Peptides are groups of consecutive amino acids in proteins, the units of LCV may be continuous or discontinuous. With gathering representative sequences for Quintessential Set that covers all motifs, flexible combinations among extracted LCVs and the units within them may facilitate extensible use of the LCV approach in prediction of big and complicated domains.

CONCLUSIONS

The LCV approach is a new strategy for variable extraction. It is effective for feature extraction from a quintessential set and for constructing a good prediction model for HTH motifs. When faced with samples of unequal length, approaches calculating sorts of properties of amino acid or making statistics information of the samples as a whole are common practice. So traditional methods show weakness owing to the overall treatment of sequences, whereas the LCV approach draws local sequence information that can be regarded as an alternative to local residue frequency information for certain residue combinations. But it is almost impossible to exhaust the same LCV variables because of the rapid exponential growth of variable count while increasing the statistical residues and interval. In contrast, out of a huge number of all possible residue combinations at different locations, those remarkable ones are singled out in LCS mining phase, which convey potential information on the quintessential set in the form of various LCSs. After refinement to LCS set, a non-redundant LCV set is formed. In the matching process, LCVs also take advantage of the discontigous information of target protein sequences. LCVs may reveal long-range interactions with biological implications according to distance analysis and the extreme helix-forming propensity. In other words, the factors contributing to the formation of HTH motif in proteins are conveyed by LCVs, and the relationship among LCVs may determine whether a protein potentially contains a HTH motif.

Generally, various quantized signals for pattern recognition are regarded as obeying a normal or Gaussian distribution, as supported by many research results. However, this does not apply to all cases. The LCV method introduces the concept that signal distribution varies in position, combination pattern and count, and may not have one central position in samples like normal distribution. This idea, in which signals are treated as a non-Gaussian-distribution, attributes the mechanism of occurrence of certain motifs (or patterns) to some multi-combination factors in sequence (i.e. the manner of residue combination and the position and count of these specific combination patterns).

Prediction results of newly reported HTH-containing proteins compared with GYM2.0 validates the prediction ability of LCV approach. Comparisons with profile-HMM models from the Pfam protein families database show that the LCV approach maintains a good balance between true positive data and negative data. Though the LCV and profile-HMM approaches are both based on sequence similarity, the LCV approach focus on those highly presented residue combinations (i.e. LCVs) from quintessential set and ignores the sequence order of LCV location, and thus makes flexible, comprehensive utilization of sequence information.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors wish to thank the two anonymous reviewers whose comments are very helpful in strengthening the presentation of this article.

FUNDING

National Natural Science Foundation of China grants [20675057, 20705024]. Funding for open access charge: National Natural Science Foundation of China.

Conflict of interest statement. None declared.

REFERENCES

1. Crick,F. (1970) Central dogma of molecular biology. *Nature*, **227**, 561–563.
2. Bussard,A.E. (2005) A scientific revolution? The prion anomaly may challenge the central dogma of molecular biology. *EMBO Rep.*, **6**, 691–694.
3. Sorokin,V., Severinov,K. and Gelfand,M.S. (2008) Systematic prediction of control proteins and their DNA binding sites. *Nucleic Acids Res.*

4. Yu,D., Chen,C. and Chen,Z. (2001) Evidence for an important role of WRKY DNA binding proteins in the regulation of NPR1 gene expression. *Plant Cell*, **13**, 1527–1540.
5. Frampton,J., Leutz,A., Gibson,T. and Graf,T. (1989) DNA-binding domain ancestry. *Nature*, **342**, 134.
6. Rosinski,J.A. and Atchley,W.R. (1999) Molecular evolution of helix-turn-helix proteins. *J. Mol. Evol.*, **49**, 301–309.
7. Aravind,L., Anantharaman,V., Balaji,S., Babu,M.M. and Iyer,L.M. (2005) The many faces of the helix-turn-helix domain: transcription regulation and beyond. *FEMS Microbiol. Rev.*, **29**, 231–262.
8. Ferrer-Costa,C., Shanahan,H.P., Jones,S. and Thornton,J.M. (2005) HTHquery: a method for detecting DNA-binding proteins with a helix-turn-helix structural motif. *Bioinformatics*, **21**, 3679–3680.
9. Shanahan,H.P., Garcia,M.A., Jones,S. and Thornton,J.M. (2004) Identifying DNA-binding proteins using structural motifs and the electrostatic potential. *Nucleic Acids Res.*, **32**, 4732–4741.
10. McLaughlin,W.A. and Berman,H.M. (2003) Statistical models for discerning protein structures containing the DNA-binding helix-turn-helix motif. *J. Mol. Biol.*, **330**, 43–55.
11. Ahmad,S., Gromiha,M.M. and Sarai,A. (2004) Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information. *Bioinformatics*, **20**, 477–486.
12. Pellegrini-Calace,M. and Thornton,J.M. (2005) Detecting DNA-binding helix-turn-helix structural motifs using sequence and structure information. *Nucleic Acids Res.*, **33**, 2129–2140.
13. Wong,Y.H., Lee,T.Y., Liang,H.K., Huang,C.M., Wang,T.Y., Yang,Y.H., Chu,C.H., Huang,H.D., Ko,M.T. and Hwang,J.K. (2007) KinasePhos 2.0: a web server for identifying protein kinase-specific phosphorylation sites based on sequences and coupling patterns. *Nucleic Acids Res.*, **35**, W588–W594.
14. Huang,J., Li,T., Chen,K. and Wu,J. (2006) An approach of encoding for prediction of splice sites using SVM. *Biochimie*, **88**, 923–929.
15. Lo,S.L., Cai,C.Z., Chen,Y.Z. and Chung,M.C. (2005) Effect of training datasets on support vector machine prediction of protein-protein interactions. *Proteomics*, **5**, 876–884.
16. Konieczny,L., Brylinski,M. and Roterman,I. (2006) Gauss-function-Based model of hydrophobicity density in proteins. *In Silico Biol.*, **6**, 15–22.
17. Finn,R.D., Tate,J., Mistry,J., Coggill,P.C., Sammut,S.J., Hotz,H.R., Ceric,G., Forslund,K., Eddy,S.R., Sonnhammer,E. *et al.* (2008) The Pfam protein families database. *Nucleic Acids Res.*, **36**, D281–D288.
18. Strope,P.K. and Moriyama,E.N. (2007) Simple alignment-free methods for protein classification: a case study from G-protein-coupled receptors. *Genomics*, **89**, 602–612.
19. Mathee,K. and Narasimhan,G. (2003) Detection of DNA-binding helix-turn-helix motifs in proteins using the pattern dictionary method. *RNA Polym. Assoc. Factors, PTC*, **370**, 250–264.
20. Brennan,R.G. and Matthews,B.W. (1989) The helix-turn-helix DNA binding motif. *J Biol Chem*, **264**, 1903–1906.
21. Schultz,J., Copley,R.R., Doerks,T., Ponting,C.P. and Bork,P. (2000) SMART: a web-based tool for the study of genetically mobile domains. *Nucleic Acids Res.*, **28**, 231–234.
22. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I. *et al.* (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.
23. Gribskov,M. and Veretnik,S. (1996) Identification of sequence patterns with profile analysis. *Computer Methods for Macromolecular Sequence Analysis*, **266**, 198–212.
24. Chang,C. and Lin,C. (2001) LIBSVM: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.
25. Bairoch,A., Consortium,U., Bougueleret,L., Altairac,S., Amendolia,V., Auchincloss,A., Argoud-Puy,G., Axelsen,K., Baratin,D., Blatter,M.C. *et al.* (2009) The Universal Protein Resource (UniProt) 2009. *Nucleic Acids Res.*, **37**, D169–D174.
26. Pace,C.N. and Scholtz,J.M. (1998) A helix propensity scale based on experimental studies of peptides and proteins. *Biophys. J.*, **75**, 422–427.
27. Cai,Y.D. and Chou,K.C. (2005) Predicting enzyme subclass by functional domain composition and pseudo amino acid composition. *J. Proteome Res.*, **4**, 967–971.
28. Liao,L. and Noble,W.S. (2003) Combining pairwise-sequence similarity and support vector machines for detecting remote protein evolutionary and structural relationships. *J. Comput. Biol.*, **10**, 857–868.
29. Cai,C.Z., Han,L.Y., Ji,Z.L. and Chen,Y.Z. (2004) Enzyme family classification by support vector machines. *Proteins Struct. Funct. Bioinform.*, **55**, 66–76.
30. Kunik,V., Meroz,Y., Solan,Z., Sandbank,B., Weingart,U., Ruppin,E. and Horn,D. (2007) Functional representation of enzymes by specific peptides. *PLOS Comput. Biol.*, **3**, 1623–1632.