Discovering protein-DNA binding sequence patterns using association rule mining

Leung, Kwong-Sak; Wong, Ka-Chun; Chan, Tak-Ming; Wong, Man-Hon; Lee, Kin-Hong; Lau, Chi-Kong; Tsui, Stephen K.W.

Published in:
Nucleic Acids Research

Published: 01/10/2010

Document Version:
Final Published version, also known as Publisher’s PDF, Publisher’s Final version or Version of Record

License:
CC BY-NC

Publication record in CityU Scholars:
Go to record

Published version (DOI):
10.1093/nar/gkq500

Publication details:
Leung, K-S., Wong, K-C., Chan, T-M., Wong, M-H., Lee, K-H., Lau, C-K., & Tsui, S. K. W. (2010). Discovering protein-DNA binding sequence patterns using association rule mining. Nucleic Acids Research, 38(19), 6324-6337. [gkq500]. https://doi.org/10.1093/nar/gkq500

Citing this paper
Please note that where the full-text provided on CityU Scholars is the Post-print version (also known as Accepted Author Manuscript, Peer-reviewed or Author Final version), it may differ from the Final Published version. When citing, ensure that you check and use the publisher’s definitive version for pagination and other details.

General rights
Copyright for the publications made accessible via the CityU Scholars portal is retained by the author(s) and/or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights. Users may not further distribute the material or use it for any profit-making activity or commercial gain.

Publisher permission
Permission for previously published items are in accordance with publisher’s copyright policies sourced from the SHERPA RoMEO database. Links to full text versions (either Published or Post-print) are only available if corresponding publishers allow open access.

Take down policy
Contact lbscholars@cityu.edu.hk if you believe that this document breaches copyright and provide us with details. We will remove access to the work immediately and investigate your claim.

Download date: 08/10/2023
Discovering protein–DNA binding sequence patterns using association rule mining

Kwong-Sak Leung¹, Ka-Chun Wong¹,*, Tak-Ming Chan¹, Man-Hon Wong¹, Kin-Hong Lee¹, Chi-Kong Lau² and Stephen K. W. Tsui²,³

¹Department of Computer Science & Engineering, ²School of Biomedical Sciences, The Chinese University of Hong Kong, and ³Hong Kong Bioinformatics Centre, Shatin, N. T., Hong Kong, China

Received February 23, 2010; Revised May 17, 2010; Accepted May 18, 2010

ABSTRACT
Protein–DNA bindings between transcription factors (TFs) and transcription factor binding sites (TFBSs) play an essential role in transcriptional regulation. Over the past decades, significant efforts have been made to study the principles for protein–DNA bindings. However, it is considered that there are no simple one-to-one rules between amino acids and nucleotides. Many methods impose complicated features beyond sequence patterns. Protein-DNA bindings are formed from associated amino acid and nucleotide sequence pairs, which determine many functional characteristics. Therefore, it is desirable to investigate associated sequence patterns between TFs and TFBSs. With increasing computational power, availability of massive experimental databases on DNA and proteins, and mature data mining techniques, we propose a framework to discover associated TF–TFBS binding sequence patterns in the most explicit and interpretable form from TRANSFAC. The framework is based on association rule mining with Apriori algorithm. The patterns found are evaluated by quantitative measurements at several levels on TRANSFAC. With further independent verifications from literatures, Protein Data Bank and homology modeling, there are strong evidences that the patterns discovered reveal real TF–TFBS bindings across different TFs and TFBSs, which can drive for further knowledge to better understand TF–TFBS bindings.

INTRODUCTION
We first introduce protein–DNA bindings in this section. Existing bioinformatics methods are briefly described, followed by the layout of this article.

PROTEIN–DNA binding
Protein–DNA binding plays a central role in genetic activities such as transcription, packaging, rearrangement, and replication (1,2). Therefor, it is very important to identify and understand the protein–DNA bindings as the basis for further deciphering biological systems. We focus on protein–DNA bindings between transcription factors (TFs) and transcription factor binding sites (TFBSs), which are the primary regulatory activities with abundant data support. TFs bind in a sequence-specific manner to TFBSs to regulate gene transcription (gene expression). The DNA binding domain(s) of a TF can recognize and bind to a collection of similar TFBSs, from which a conserved pattern called motif can be obtained. TFBSs, the nucleotide fragments bound by TFs, are usually short (usually about 5–20 bp) in the cis-regulatory/intergenic regions and can assume very different locations from the transcription start site.

It is expensive and laborious to experimentally identify TF–TFBS binding sequence pairs, for example, using DNA footprinting (3) or gel electrophoresis (4). The technology of chromatin immunoprecipitation (ChIP) (5,6) measures the binding of a particular TF to DNA of co-regulated genes on a genome-wide scale in vivo, but at low resolution. Further processing are needed to extract precise TFBSs (7). TRANSFAC (8) is one of the largest and most representative databases for regulatory elements including TFs, TFBSs, nucleotide distribution matrices of the TFBSs and regulated genes. The data are expertly annotated and manually curated from peer-reviewed and experimentally proved publications. Other annotation databases of TF families and binding domains are also available [e.g. PROSITE (9), Pfam (10)]. It is even more difficult and time-consuming to extract high-resolution 3D TF–TFBS complex structures with X-ray crystallography or nuclear magnetic resonance (NMR) spectroscopic analysis. Nevertheless, the high-quality TF–TFBS binding structures provide valuable insights into verifications of putative principles of
Existing methods

The first attempt of computational methods related to TF–TFBS bindings was to discover the motifs of TF domains and TFBSs separately. Moreover, researchers have been trying hard to generalize the one-to-one binding codes from existing 3D structures. Data mining methods have also been proposed with feature transformations and machine learning to decipher complicated binding rules. They are briefly described as follows:

Motif discovery. TF domains and TFBSs are somewhat conserved due to their functional similarity and importance. By exploiting conservation in the sequences, Bioinformatics methods called motif discovery save some of the expensive and laborious laboratory experiments. Motif discovery (6) can be categorized into two types: (i) motif matching and (ii) de novo motif discovery. (i) Motif matching is to identify putative TF domains (9,10) or TFBSs (12) based on motif knowledge obtained from annotated data. (ii) de novo motif discovery predicts conserved patterns without knowledge on their appearances, based on certain motif models and scoring functions (13,14) from a set of protein/DNA promoter sequences with similar regulatory functions. While de novo motif discovery is successful for well-conserved TF functional domain motifs, the counterpart for TFBSs remains very challenging with poor performances on real benchmarks (6,15,16). A significant limitation of motif discovery is the lack of linkage between the binding counterparts for revealing TF–TFBS relationships.

One-to-one binding codes. Numerous studies have been carried out to analyze existing protein–DNA binding 3D structures comprehensively (2,17,18) or with focus on specific families (1) [e.g. zinc fingers (19)]. Various properties have been discovered concerning, e.g. bonding and force types, TF conservation and mutation (1), and bending of the DNA (17). Some are already applicable to predict binding amino acids on the TF side (20). However, annotated data are far from complete. Alternatively, researchers have sought hard for general binding ‘codes’ between proteins and DNA, in particular the one-to-one mapping between amino acids from TFs and nucleotides from TFBSs. Despite many proposed one-one binding propensity mappings (1,21,22), it has come to a consensus that there are no simple binding ‘codes’ (23).

Data mining. In the hope of better understanding for protein–DNA bindings, many data mining approaches have also been proposed (24). Researchers employ and transfer additional detailed information such as base compositions, structures, thermodynamic properties (25,26) as well as expressions (27), into sophisticated features to fit into certain data mining techniques. Although some approaches may provide interpretable rules, most of them have stringent data requirements which cannot be obtained trivially. Existing data beyond sequences are also insufficient and limited for practitioners. These methods usually extract complicated features rather than working on interpretable data directly. Many data mining techniques, such as neural networks, support vector machines (SVM) (28) and regressions (24), may generate rules which are not trivial to interpret. Furthermore, many data mining approaches are based on specific families or particular data sets, where the generality of the results are limited. On the other hand, sequences serve as the most handy primary data that carry important information for protein–DNA bindings (23). It is desirable to make use of the large-scale and comprehensive sequence data to mine explicit and interpretable TF–TFBS binding rules.

Article layout

In this article, we propose a framework based on association rule mining to discover protein–DNA binding sequence patterns from TRANSFAC. The article layout is as follows: the proposed methods are presented in the next section: ‘Materials and Methods’ section; experimental results and verifications are reported in sections ‘Results and Analysis’ and ‘Verifications’ section, respectively; and finally we have the ‘Discussion’ section for the approach.

MATERIALS AND METHODS

In this section, we propose a framework for mining, discovering and verifying TF–TFBS bindings on large-scale databases. The framework starts from data cleansing and transformation on TRANSFAC, and then applies association rule mining to discover TF-TFBS binding sequence patterns. Comprehensive 3D verifications and evaluations are carried out on PDB. Detailed bonding analysis is performed to provide strong support to the discovered rules.

In the following subsections, Apriori algorithm for association rule mining is first introduced. We then elaborate how the algorithm is applied to protein–DNA binding pattern discovery. Finally, we present how the data are preprocessed for the task with a running example.

Association rule mining and Apriori Algorithm

Association rule mining (29) aims at discovering frequently co-occurring items, called frequent itemsets, from a large number of data samples above a certain count threshold (minimum support) (30). The support of an itemset is defined as the number of data samples where all the items in the itemset co-occur. In the case of protein–DNA binding, the binding domains of TFs can recognize and form strong bindings with certain sequence-specific patterns of the TFBSs. Therefore, they are likely to co-occur frequently among the combinations between all
possible TF and TFBS subsequences, and can be thus identified by association rule mining. In this study, we use the notation of $k$-mer (a subsequence with $k$ amino acid or nucleotide residues) to represent a candidate item. A frequent TF–TFBS itemset is a TF $k$-mer and TFBS $k$-mer (the two $k$’s can be different) pair, or simply a pair, co-occurring with a frequency no less than the minimum support in the TF–TFBS sequence records (TRANSFAC database).

Apriori algorithm proposed by Agrawal et al. (29) is a classical approach to find out frequent itemsets. It is outlined in Algorithm 1 in the Appendix 1. It is a branch and bound algorithm for discovering association rules in a database. With its downward closure property, an optimal performance is guaranteed. The algorithm first obtains frequent 1-itemsets. Iteratively, it uses the frequent itemsets to generate all possible candidate $(n + 1)$-itemsets. They are then evaluated for their support (30). If the support of an $(n + 1)$-itemset is lower than a threshold, the $(n + 1)$-itemset is removed. After the removal, the resultant $(n + 1)$-itemsets are the frequent $(n + 1)$-itemsets. The above procedure is repeated until an empty set is found.

Discovering associated TF–TFBS sequence patterns

To formulate the TF–TFBS sequence pattern discovery problem into association rule mining, we have to transform the protein–DNA binding records into the formats of itemsets ($k$-mers). An illustrative example for the TF–TFBS binding records from TRANSFAC 2008.3 is shown in Figure 1. The TF (e.g. T01333 RXR-γ) can bind to several TFBS DNA sequences. The DNA sequences may be different in lengths due to experimental methods and noises. Both the TF and TFBS sequences are chopped into overlapping short $k$-mers, as illustrated in Figure 2 (first part). They together with the corresponding reverse complements (e.g. GACCT and reverse complement: AG GTC) form one data sample. To generate the itemsets, all the $k$-mers are recorded in a binary array where appearing $k$-mer is marked 1; and 0 otherwise. Thus, the length of the array depends on the number of all possible TF $k$-mers and TFBS $k$-mers (Figure 2, second part). Since $k$ is usually short (4–6), all the possible $4^k$ combinations of TFBS DNA $k$-mers can be adopted. However, it is computationally infeasible to obtain all the possible $20^k$ combinations of TF $k$-mers. Thus a data-driven approach is employed by scanning the whole TRANSFAC to obtain frequent TF amino acid $k$-mers.

Since there are multiple TFBSs for each TF (e.g. Figure 1), a question arises: how to define the ‘commonly found’ TFBS $k$-mers of a TF? Without loss of generality, the majority rule (31) is applied. If the majority of a TF’s TFBS sequences contains a certain DNA residue $k$-mer, then the $k$-mer is considered ‘commonly found’. We set the majority to be 50% for TFBS $k$-mers. We only count the number of TFBS sequences in which a certain $k$-mer appears, in order not to be biased by multiple occurrences of the $k$-mer appearing in only a few TFBS sequences. Figure 1 illustrates an example where there are five TFBS sequences. The TFBS DNA $k$-mer AGGTC (or its reverse complement: GACCT) can be found in three of the TFBS sequences. The $k$-mer appears in 60% (3/5) of the TFBS sequences of the TF, and thus is considered ‘commonly found’. On the other hand, GTTCA is not considered ‘commonly found’ because it only appears in 2 (40%) out of the 5 TFBS sequences of the TF.

After all valid TF data samples are transformed into itemsets, Apriori algorithm is applied to generate frequent TF–TFBS $k$-mer sequence patterns (the links in Figure 2, second part). The special feature in this study is that the co-occurring pairs should contain both TF and TFBS $k$-mers, as illustrated in the third part of Figure 2. In the current study, we only consider one TF $k$-mer with one TF BS $k$-mer in the frequent itemsets, but it is straightforward to generalize it to be multiple TF and TFBS $k$-mers in principle. The huge computational intensity for the generalization, when applied on the large TRANSFAC database, prevents us from doing so at this time. Finally, the association rules are computed based on the confidence measurements for the frequent itemsets, which are defined as follows:

$$
\text{conf}(k\text{-mer}_{\text{DNA}} \Rightarrow k\text{-mer}_{\text{AA}}) = \frac{\text{support}(k\text{-mer}_{\text{DNA}} \cap k\text{-mer}_{\text{AA}})}{\text{support}(k\text{-mer}_{\text{DNA}})}
$$

$$
\text{conf}(k\text{-mer}_{\text{DNA}} \Leftarrow k\text{-mer}_{\text{AA}}) = \frac{\text{support}(k\text{-mer}_{\text{DNA}} \cap k\text{-mer}_{\text{AA}})}{\text{support}(k\text{-mer}_{\text{AA}})}
$$

where $\text{conf}(k\text{-mer}_{\text{DNA}} \Rightarrow k\text{-mer}_{\text{AA}})$ is called forward confidence, $\text{conf}(k\text{-mer}_{\text{DNA}} \Leftarrow k\text{-mer}_{\text{AA}})$ is called backward confidence and support($(X)$ is the support of itemset $X$. For each association rule, its forward confidence measures the posterior probability that the corresponding amino acid $k$-mer can be found in a TF’s sequence if the DNA $k$-mer is commonly found in the TF’s TFBS sequences. Its backward confidence measures the posterior probability that the corresponding DNA $k$-mer can be commonly found in a TF’s TFBS sequences if the amino acid $k$-mer is found in the TF’s sequence. The minimum of them is taken as confidence in this article. The higher the confidence, the better the association rule is (Figure 2, fourth part). The whole proposed approach is summarized in Figure 2.

Data preparation

To apply the methodology on TRANSFAC, TF and TFBS data were downloaded and extracted from the flat
files of TRANSFAC 2008.3 [a free public (older) version is also available (http://www.gene-regulation.com/pub/database.html)]. The entries without sequence data were discarded. Since a TF can bind to one or more TFBSs, TFBS data were grouped by TF. TFBS sequences were extracted for each TF to form a TF data set—a TF sequence and the corresponding TFBS sequences—and finally to be transformed into itemsets. To avoid sampling error, TF data sets with less than five TFBS sequences were discarded. Furthermore, the redundancy of TF sequences was removed by BLASTClust using 90% TF sequence identity (32). Only one TF data set was selected for each cluster. Note that we only used sequence data in TRANSFAC. None of the prior information (e.g. the binding domains of TFs) other than sequences was used. Importantly, it turns out that the results

Figure 2. Flowchart of the proposed framework to discover association rules from TRANSFAC.
RESULTS AND ANALYSIS

In this section, the discovered rules are reported, followed by analysis with different measurements.

Rules discovered

Varying \( k \) from 4 to 6 for both TF \( k \)-mers and TFBS \( k \)-mers, we have obtained nine sets of associated pairs. For each set of pairs, the forward and backward confidences of each pair were calculated. Then, the pairs in the same set were sorted by the minima of their forward and backward confidences in descending order. The nine sets of rules (pairs) exhibit a similar trend that the number of rules (pairs) decreases as the association criterion becomes more stringent (with higher confidence levels). The TFBS 5-mers settings in general show the most available rules when the confidence level is high (\( \geq 0.5 \)), indicating more conserved and significant results. Therefore, we focus on them and use TFBS 5-mer–TF 5-mer as the representative example throughout the article. The results for all other settings are available in the Supplementary Data.

The number of rules (pairs) discovered is summarized in Table 1. For instance, there are 70 TF 5-mer–TFBS 5-mer pairs without any further removal (in the \( N \) column) with both forward and backward confidences \( \geq 0.5 \). Considering direct and reverse complement TFBS DNA \( k \)-mers as equivalent, we further removed the duplicated pairs (e.g. leaving AGGTC–CGCTG and removing GAC CT–CGCTG because AGGTC and GACCT are reverse complements). The results are shown in the \( N \) column in Table 1. For instance, the 70 TF 5-mer–TFBS 5-mer pairs were reduced to 35 at a confidence level of 0.5. Furthermore, we found that most pairs could be merged together to form a longer pair. For instance, GGTCAGTT and GGTCAGTTA could be merged to form a pair GGTCAGTTA. Thus the pairs have been merged and the rule numbers are shown in the \( N_m \) column in Table 1. For instance, 35 TF 5-mer–TFBS 5-mer pairs are merged to form 11 merged pairs when the confidence level is equal to 0.5.

Quantitative analysis

To evaluate the number of TF data sets supporting each pair (support), the support for each pair was counted. In general, more supports are found when the confidence level is increased. For instance, the average support of the TFBS 5-mer–TF 5-mer pairs is generally increased when the confidence level is increased in the \( S \) column of Table 1. The overall results are summarized in Supplementary Table S4.

Support is considered the degree of co-occurrence between a TF amino acid \( k \)-mer and a TFBS DNA \( k \)-mer. Forward and backward confidences consider the cases when either one of them is absent. Some may have questions about the remaining case. How about the case when both of them are absent? To take the case into account, \( \phi \)-coefficients (35) were measured for each pair, as shown in the \( \phi \) column in Table 2. The overall results are summarized in Supplementary Table S5. Most values are \( >0.4 \), indicating that positive correlations exist among pairs.

Consider the following scenario: if a TFBS DNA \( k \)-mer and a TF amino acid \( k \)-mer are both frequently found in the data sets, it will be very likely that they co-occur...
frequently merely by chance. To tackle such scenario, forward and backward confidences do play their important roles in pruning them. But for clarity, lift (36) that estimates the ratio of the actual support to the expected support was measured for each pair, where the expected support was calculated from the random model that the TFBS DNA k-mer is independent of the TF amino acid k-mer for each pair. For instance, the average lift for the TFBS 5-mer–TF 5-mer pairs is shown in the L column in Table 2. The overall results are summarized in Supplementary Table S6. Most values of the lift are >5. Thus the DNA residue k-mer and the amino acid residue k-mer of most pairs co-occur at least five times more frequently than the prediction based on the independent assumption made by the lift measurement.

To estimate the validity of the pairs, both forward and backward convictions (the same directions as the forward and backward confidences, respectively) (36) were measured for each pair. The measurements were averaged for each set of pairs. For instance, the average forward and backward convictions for the TFBS 5-mer–TF 5-mer pairs is shown in the FC and BC columns in Table 2. The overall results are summarized in Supplementary Tables S7 and S8. Most values are >1. The pairs commit fewer errors than the prediction based on the statistically independent assumption made by the measurements: forward and backward convictions. In other words, the pairs would have committed more errors if the association between its TFBS k-mer and TF k-mer had happened purely by chance.

Annotation analysis

If the pairs in our results are the actual binding cores between TFs and TFBSs, most of their TF amino acid k-mers should be inside DNA binding domains. Thus, the TF amino acid k-mers were scanned in TRANSFAC to check whether they were within the annotated DNA binding domains. As stated in the previous section, the set of TFBS 4-mer–TF 4-mer pairs constitutes all the pairs in the other sets by the downward closure property. Thus only the TF amino acid 4-mers of the set of TFBS 4-mer–TF 4-mer pairs were needed for the checking: of the 792 TF amino acid 4-mers, 92.2 % of them were found within the DNA binding domains listed in the ‘PFAM 18’ list downloaded from DBD (37) on 25 January 2010.

Empirical analysis

Since the numbers of results are quite large, they are tabulated in a statistical perspective in the previous sections. This section provides readers with empirical insights into the results obtained. Comparing with the other sets, the set of TFBS 5-mer–TF 5-mer pairs shows its relative invariability to confidence level pruning. Thus, it motivates us to have an in-depth empirical analysis on them. They are listed in Table 3.

Among the 131 pairs in Table 3, the TFBS DNA k-mers are quite conserved. There are only 15 distinct TFBS DNA k-mers. Each TFBS DNA k-mer forms pairs with 8.73 TF amino acid k-mers on average. One of the reasons may be the specificity of DNA residue, is lower in view of its alphabet size (4) as compared to the amino acid alphabet size (20).

To act as a DNA binding protein, a TF needs to provide a basic interacting surface for the recognition of major/ minor grooves as well as the phosphate backbone of DNA. Therefore, we searched through the set of pairs in Table 3 to count the occurring frequency for each residue. Interestingly, we found that the basic residues, lysine (50 times) and arginine (131 times), occur at the highest frequency among 131 pairs of TFBS–TF. On the other hand, the hydrophobic residues (38) such as isoleucine (15) and valine (13) occur at the lowest frequency. These results suggest the potential of the TF sequences for being the binding sequences between TFs and TFBSs. On the other hand, as the nucleotides of TFBSs are somehow negatively charged, it can be deduced that their binding amino acid residues of TFs should be positively charged. Thus the occurring frequencies were further examined. Among the 131 pairs, the positively charged residues: arginine (R) and lysine (K) occur 131 and 50 times, respectively. In contrast, the negatively charged residues aspartic acid (D) and glutamic acid (E) occur 8 and 30 times, respectively. Such discrepancy supports their potential for being the binding sequences between TFs and TFBSs.

Experimental analysis

This section follows the same approach in empirical analysis. The set of TFBS 5-mer–TF5mer pairs in Table 3 is selected for experimental analysis. Out of the 131 pairs, 5 of them were selected and analyzed. The first pair is GGTCA–CEGCK, which have been experimentally proved as binding sequences in Ref. (39). The TF amino acid 5-mer (CEGCK) is considered part of P-box (CEGCKG) within the DNA binding domain of Bp-nhr-2, which is believed to bind the DNA k-mer (G GTCA). The second pair is AAACA–IRHNL mentioned in Ref. (40). Based on the corresponding PDB entry 3CO6, it is believed that the pair was the binding pair between a TF and a TFBS as shown in Figure 3. Similarly, the remaining pairs are GATAA–NACGL, G GTCA–GFFRR and CTCCC–LRYYY. They are found as binding pairs in PDB entries 3DFV (41), 3DZY (42) and 2NNY (43) as shown in Figure 3a, b and c, respectively. The above five pairs reveal that the pairs generated from the proposed approach have biological evidences in literatures. Among the previous figures, two of them (3CO6 and 2NNY) were further analyzed in terms of hydrogen bonding, which also means the specificity of the interaction between amino acids and the bases, as shown in Figure 4a and b. We have also highlighted the hydrogen bonds as black lines as well as the residues that make contact with the base (only predicted residues), which are the evidence of the significance and accuracy of the prediction of the TF–TFBS pairs. Nevertheless, as the proposed approach is applied on a large-scale database, such extensive and detailed analysis of
all the binding core pairs discovered are not practical. Therefore, a scalable verification approach will be presented in the next section to verify the massive results generated.

**VERIFICATIONS**

In this section, we try to verify the discovered pairs with external data sources, in particular the 3D protein-DNA complex structures experimentally determined from PDB.
Homology modeling has also been done for further verifications.

Verification by PDB

In this article, PDB is selected for providing 3D protein–DNA complex data for 3D structural verification. The PDB data were downloaded from RCSB PDB (http://www.pdb.org) from 16 September 2009 to 22 September 2009, where the protein–DNA complexes were selected based on the entry-type list provided in ftp://ftp.wwpdb.org/.

For each set of pairs in Supplementary Table S2, each pair is independently evaluated as shown in Figure 5. For each pair, its TF $k$-mer is used to query which PDB chain has the TF $k$-mer. Once the corresponding set of PDB chains has been identified and returned, its redundancy is removed by BLASTClust using 90% sequence identity (32). The removal is to ensure that redundant PDB chains are not double counted. After the removal, the pair is evaluated for binding in the 3D space:

- A TFBS $k$-mer–TF $k$-mer pair is considered binding for a PDB chain if and only if an atom of the TFBS $k$-mer and an atom of the TF $k$-mer are close to each other. Two atoms are considered close if and only if their distance is $<3.5$ Å (25,28).

With the pair evaluated in its PDB chains, its PDB chains can be classified into the following three categories:

- PDB chains only having the TF $k$-mer ($a$)
- PDB chains having both TF $k$-mer and TFBS $k$-mer
  - The pair binds together ($b$)
  - The pair does not bind together ($c$)

Thus the number of chains in each category is counted and converted into the following performance metrics:

- TFBS prediction score = $(b+c)/(a+b+c)$
- TFBS binding prediction score = $b/(a+b+c)$
- Binding prediction score = $b/(b+c)$

Given the resultant PDB chains queried by a TF $k$-mer, TFBS prediction score measures the proportion of PDB chains that contain the corresponding TFBS $k$-mer.

In other words, it measures the backward confidence of a pair in PDB. TFBS binding prediction score is a more stringent metric. It measures the proportion of PDB chains that have the corresponding TFBS $k$-mer binding with the queried TF $k$-mer. Lastly, binding prediction score is the most important metric. It measures the proportion of PDB chains in which the pair is really binding. To verify the cases when $(b+c) = 0$ (i.e. the pairs do not appear in PDB), homology modeling is also performed.

For each setting, we have a set of pairs. For each pair, the above performance metrics are calculated. The overall results are averaged and summarized in Supplementary Tables S9–S11. For each setting, we also have a set of merged pairs. For each merged pair, the above performance metrics are also calculated. The overall results are averaged and summarized in Supplementary Tables S12–S14. Note that the most conservative calculation has been used for each performance metric for each pair. If a
performance metric of a pair does not have enough PDB data for calculation, a value of zero will be given to the performance metric of the pair. For instance, the cases when \((b + c) = 0\) or \((a + b + c) = 0\). Despite the above setting, the performance metrics of the pairs still have reasonable performances. They are shown to be significantly better than the maximal performance of 50 random runs in a later section.

Nevertheless, although the above metrics can capture the performance of a pair quantitatively, the most important point is to know how many generated pairs could be verified [with at least one binding evidence in PDB data \((b > 0)\)]. To gain more insights, the number of pairs with at least one related PDB chain \([b+c > 0]\) are tabulated in Supplementary Tables S15 and S16. Correspondingly, the percentage of verified pairs \(\left(\frac{\text{Number of pairs with } b > 0}{\text{Number of pairs with } (b+c) > 0}\right)\) are calculated and tabulated in Supplementary Tables S17 and S18. In the tables, the percentage of verified pairs is high enough to justify that the proposed approach has produced pairs proven to be binding in PDB. For instance, the statistics for the TFBS 5-mer–TF 5-mer pairs is extracted in Table 4 and Figure 6. Among the 80 TFBS 5-mer–TF 5-mer pairs with at least one related PDB chain \([b+c > 0]\) when the confidence level = 0.0, more than 81% of them have at least one binding evidence \((b > 0)\).

The TFBS–TF pairs that we found to have binding evidences in the PDB show typical structural features of DNA–protein interactions. Such features include the ‘recognition helix’ of the DNA–binding protein making base contacts in the major groove and direct hydrogen bonds between the side chains and the bases. These interactions play the crucial role in the DNA recognition and site-specific binding, respectively (44). Interestingly, the nucleotides of TFBS are located in the major groove of the DNA, which are close to, and make contacts with the amino acids of the ‘recognition helix’ of the TF (as for example shown in Figure 3).

The verification is considered satisfactory since those pairs not found in PDB \([b+c = 0]\) may be unannotated discovery as shown in the following verification by homology modeling.

**Verification by homology modeling**

Regarding the pairs without any related PDB chain \([b+c = 0]\), there is no PDB data for us to verify them. Thus, we have taken the most conservative approach to
| ID       | Factor name                        | ID       | Factor name                        |
|----------|------------------------------------|----------|------------------------------------|
| T0003    | AS-CT3                             | T00842   | Tra-1 (long form)                  |
| T0004    | Adf-1                              | T01950   | HNF-1-z-B                          |
| T0013    | DDX1                               | T04378   | Mad                                |
| T0019    | AhR                                | T08676   | STAT6                              |
| T0026    | Antp                               | T00887   | ARF1-isoform1                      |
| T0063    | Bcd                                | T01085   | abaA                               |
| T0077    | CACCC-binding factor               | T01112   | EBF1-L                             |
| T0079    | Cad                                | T02422   | HNF-4                              |
| T0080    | CBFI                               | T04446   | Nkx-5-1                            |
| T0104    | C/EBP x                           | T08787   | ARF1-isoform1                      |
| T0106    | C/EBP                              | T01922   | EAR2                               |
| T0108    | C/EBP                              | T09787   | Dr-1A                              |
| T0111    | c-Ets-2                            | T00887   | Opaque-2                           |
| T0112    | c-Ets-1                            | T09791   | GSTA-2                             |
| T0033    | Ap-2                               | T02423   | RFX1                               |
| T0063    | Bcd                                | T01092   | PITX2A                             |
| T0077    | CACCC-binding factor               | T01082   | SRY                                |
| T0079    | Cad                                | T00899   | WT1                                |
| T0080    | CBFI                               | T00910   | YB-1                               |
| T0104    | C/EBP x                           | T00915   | YY1                                |
| T0106    | C/EBP                              | T00951   | YBX1                               |
| T0108    | C/EBP                              | T00863   | Ubx                                |
| T0109    | CRE-BP                            | T00912   | abd-A                              |
| T0117    | C/EBP                              | T02422   | HNF-1-z-B                          |
| T0079    | Cad                                | T00899   | WT1                                |
| T0080    | CBFI                               | T00910   | YB-1                               |
| T0104    | C/EBP x                           | T00915   | YY1                                |
| T0106    | C/EBP                              | T00951   | YBX1                               |
| T0108    | C/EBP                              | T00863   | Ubx                                |
| T0109    | CRE-BP                            | T00912   | abd-A                              |
| T0117    | C/EBP                              | T02422   | HNF-1-z-B                          |
| T0079    | Cad                                | T00899   | WT1                                |
| T0080    | CBFI                               | T00910   | YB-1                               |
| T0104    | C/EBP x                           | T00915   | YY1                                |
| T0106    | C/EBP                              | T00951   | YBX1                               |
| T0108    | C/EBP                              | T00863   | Ubx                                |
| T0109    | CRE-BP                            | T00912   | abd-A                              |
| T0117    | C/EBP                              | T02422   | HNF-1-z-B                          |
| T0079    | Cad                                | T00899   | WT1                                |
| T0080    | CBFI                               | T00910   | YB-1                               |
| T0104    | C/EBP x                           | T00915   | YY1                                |
| T0106    | C/EBP                              | T00951   | YBX1                               |
| T0108    | C/EBP                              | T00863   | Ubx                                |
| T0109    | CRE-BP                            | T00912   | abd-A                              |
| T0117    | C/EBP                              | T02422   | HNF-1-z-B                          |
| T0079    | Cad                                | T00899   | WT1                                |
| T0080    | CBFI                               | T00910   | YB-1                               |
| T0104    | C/EBP x                           | T00915   | YY1                                |
| T0106    | C/EBP                              | T00951   | YBX1                               |
| T0108    | C/EBP                              | T00863   | Ubx                                |
| T0109    | CRE-BP                            | T00912   | abd-A                              |
| T0117    | C/EBP                              | T02422   | HNF-1-z-B                          |

Table 5. 631 TRANSFAC 2008.3 ID and factors used in this article

| ID       | Factor name                        | ID       | Factor name                        |
|----------|------------------------------------|----------|------------------------------------|
| T0003    | AS-CT3                             | T00842   | Tra-1 (long form)                  |
| T0004    | Adf-1                              | T01950   | HNF-1-z-B                          |
| T0013    | DDX1                               | T04446   | Nkx-5-1                            |
| T0019    | AhR                                | T08787   | ARF1-isoform1                      |
| T0026    | Antp                               | T01972   | REST-form2                         |
| T0063    | Bcd                                | T04539   | CBF1                               |
| T0077    | CACCC-binding factor               | T01992   | Abd-A                              |
| T0079    | Cad                                | T04610   | SRX                                |
| T0080    | CBFI                               | T08805   | WRKY1                              |
| T0104    | C/EBP x                           | T01903   | POUF3                              |
| T0106    | C/EBP                              | T01923   | CRE-z-B                            |
| T0108    | C/EBP                              | T02422   | HNF-1-z-B                          |
| T0109    | CRE-BP                            | T04539   | CBF1                               |
| T0117    | C/EBP                              | T04539   | CBF1                               |
| T0079    | Cad                                | T08823   | E2F                                |
| T0080    | CBFI                               | T08853   | myogenin                           |
| T0104    | C/EBP x                           | T08863   | S8                                 |
| T0106    | C/EBP                              | T01444   | Gfi1b                              |
| T0108    | C/EBP                              | T01085   | apaA                               |
| T0109    | CRE-BP                            | T00301   | Pax-2.1                            |
| T0117    | C/EBP                              | T00886   | CTGF                               |
| T0079    | Cad                                | T01082   | PITX2A                             |
| T0080    | CBFI                               | T01082   | PITX2A                             |
| T0104    | C/EBP x                           | T01082   | PITX2A                             |
| T0106    | C/EBP                              | T01082   | PITX2A                             |
| T0108    | C/EBP                              | T01082   | PITX2A                             |
| T0109    | CRE-BP                            | T01082   | PITX2A                             |
| T0117    | C/EBP                              | T01082   | PITX2A                             |
| T0079    | Cad                                | T01082   | PITX2A                             |
| T0080    | CBFI                               | T01082   | PITX2A                             |
| T0104    | C/EBP x                           | T01082   | PITX2A                             |
| T0106    | C/EBP                              | T01082   | PITX2A                             |
| T0108    | C/EBP                              | T01082   | PITX2A                             |
| T0109    | CRE-BP                            | T01082   | PITX2A                             |
| T0117    | C/EBP                              | T01082   | PITX2A                             |

Downloaded from https://academic.oup.com/nar/article/38/19/6324/2409524 by City University of Hong Kong Library user on 22 December 2021
assign zero to their performance metrics in the aforementioned evaluations. Nevertheless, we believe that most of those pairs are true and our approach can be used as an effective protein–DNA binding discovery tool. Thus 6 TFBS 5-mer–TF 5-mer pairs were taken and merged. The resultant pair ACCTG-SNRESARR was analyzed by homology modeling as follows:

The model of DNA–protein complex was built by homology modeling (INSIGHT II, MSI) based on the structure of the GCN4–DNA complex (1YSA) (45). Briefly, three amino acids (R234S, T236R and A238S) and two nucleotides (T29C and A31T) were mutated in the original structure. The side chains of the mutated amino acids were chosen from the rotamer database and the original structure. The side chains of the mutated and two nucleotides (T29C and A31T) were mutated in the homology modeling (INSIGHT II, MSI) based on the structure of the GCN4–DNA complex (1YSA) (45). As shown in the future.

As shown in Figure 3, we found that the pair ACCTG-SNRESARR exists in plant as the basic leucine-zipper (bZIP) transcription factor which binds to G-box binding factors (GBF) of DNA (46). Moreover, the ACCTG sequence is the consensus sequence, which is defined as G-box core and locates at the major groove of the double-stranded DNA. It is believed that the G-box core is the DNA sequence of GBF that provides the specificity of the binding to bZIP proteins. In order to further understand the interactions between the TF–TFBS, we built a model by using homology modeling based on the structure of GCN4–DNA (1YSA) complex (45). As shown in the model, the protein helix fits into the major groove of the DNA very well and forms extensive interactions (black lines) between the amino acids and the nucleotides. Interestingly, the mutations of the protein (R234S, T236R and A238S) as well as nucleotides (T29C and A31T) increases the number of hydrogen bonds compared with the original structure (1YSA), suggesting the binding specificity between this pair of TF–TFBS. In conclusion, we believe that the protein–DNA binding sequence patterns found using association rule mining on the large-scale database reveal real TF–TFBS pairs in physiologically relevant situation and this method could guide us to discover new and undescribed TF–TFBS pairs in the future.

Verification by random analysis

For each set of pairs in Supplementary Table S1, we use a random process to generate a random set with the same size and perform the same analysis as in the previous section. There are five possible outcomes of each run: true positive, false positive, true negative, false negative, and accuracy. We count the number of true positives and true negatives in the random sets and compare them with the original set to evaluate the performance of our approach. In this way, we can verify the significance of our approach in identifying new TF–TFBS pairs.

| ID   | Factor name | ID   | Factor name |
|------|-------------|------|-------------|
| T00500 | MCM1 | T0059 | MG1 |
| T01528 | RORx2 | T03178 | SQUA |
| T0309 | TFEB-isoform1 | T08300 | ER-αL |
| T01558 | RELA-p65 | T09507 | Sox-bbx1 |
| T01651 | Pax-1 | T0862 | Sox9 |
| T00789 | Tll | T01867 | Oct-2.6 |
| T00594 | RAR-α | T08630 | CAR |
| T00830 | TGA1b | T01882 | unc-86 |
| T00627 | NIT2 | T08619 | WEREWOLF |
| T01841 | WT1-del2 | T08621 | Sox9 |
| T00199 | TGF | T01866 | Oct-2.4 |
| T03975 | SPF1 | T08680 | Kdd |
| T01838 | SOX10 | T0184 | Oct-2.4 |
| T00689 | PHO2 | T08690 | Kdd |
| T01823 | Pax-2 | T08667 | SF1-1 |
| T09700 | SRY-α | T01814 | Pax-6/Pd-5a |
| T00624 | POUZ2 | T01799 | Tal-1 |
| T09718 | MSA-R | T01692 | T3R-β1 |
| T03447 | Lhx3 | T01710 | NCOA |
| T00798 | TBP | T09707 | cel-miR-84 |
| T09588 | RXR-β | T01757 | HOXD10 |
| T09649 | POU2F1b | T03975 | SPF1 |
| T00763 | SRF | T09701 | cel-miR-84 |
| T01827 | OCTFB | T09718 | hsa-miR-23a |
| T00700 | qa-1F | T09727 | hsa-miR-103 |
| T01844 | WT1-del2 | T09729 | hsa-miR-107 |
| T00710 | RAP1 | T09730 | hsa-miR-107 |
| T03717 | ZAP1 | T09723 | hsa-miR-107 |
| T00719 | RAR-α | T09731 | hsa-miR-22a |
| T01847 | AR | T09728 | hsa-miR-22a |
| T00725 | SGF-3 | T09729 | hsa-miR-22a |
| T0179 | HOXD10 | T09728 | hsa-miR-22a |
| T00737 | SAP-1a | T09729 | hsa-miR-22a |
| T03719 | ZAP1 | T09734 | hsa-miR-22a |
| T00746 | TAF-3 | T09729 | hsa-miR-22a |
| T01719 | HOXD10 | T09735 | hsa-miR-22a |
| T00751 | Sn | T09729 | hsa-miR-22a |
| T01838 | OCTFB | T09735 | hsa-miR-22a |
| T09588 | RXR-β | T09736 | hsa-miR-22a |
| T00535 | NF-1 | T03975 | SPF1 |
| T09507 | Sox-bbx1 | T09737 | hsa-miR-22a |

For each set of pairs in Supplementary Table S1, we use a random process to generate a random set with the same size and perform the same analysis as in the previous section. There are five possible outcomes of each run: true positive, false positive, true negative, false negative, and accuracy. We count the number of true positives and true negatives in the random sets and compare them with the original set to evaluate the performance of our approach. In this way, we can verify the significance of our approach in identifying new TF–TFBS pairs.

Verification by random analysis

For each set of pairs in Supplementary Table S1, we use a random process to generate a random set with the same size and perform the same analysis as in the previous section. There are five possible outcomes of each run: true positive, false positive, true negative, false negative, and accuracy. We count the number of true positives and true negatives in the random sets and compare them with the original set to evaluate the performance of our approach. In this way, we can verify the significance of our approach in identifying new TF–TFBS pairs.
number of pairs. Within a random set, its pairs were randomly sampled from all the combinations of the $k$-mers used in the proposed approach. Fifty random runs were performed. The maximal performance metrics of the 50 random runs are summarized in Supplementary Tables S19–S21. In a comparison to the proposed approach, their performance has been depicted in Figure 8. It can be observed that the performance of the proposed approach is significantly better than the best one of the 50 random runs. For instance, the binding prediction score of the 131 TFBS 5-mer–TF 5-mer pairs generated is $0.36 \pm 0.39$ on average, whereas the maximal binding prediction score over 50 random runs is only $0.00509 \pm 0.06492$ on average. Similar observation can also be drawn for their merged pairs in Supplementary Tables S22–S24. It can be concluded that the performance of the proposed approach is very unlikely to happen purely by chance in PDB.

**DISCUSSION**

In this article, we have proposed a framework based on association rule mining with Apriori algorithm to discover associated TF–TFBS binding sequence patterns in the most explicit and interpretable form from TRANSFAC. With downward closure property, the algorithm guarantees the exact and optimal performance to generate all frequent TFBS $k$-mer TF $k$-mer pairs from TRANSFAC. The approach relies merely on sequence information without any prior knowledge in TF binding domains or protein–DNA 3D structure data. From comprehensive evaluations, statistics of the discovered patterns are shown to reflect meaningful binding characteristics. According to external literatures, PDB data and homology modeling, a good number of TF–TFBS binding patterns discovered have been verified by experiments and annotations. They exhibit atomic-level bindings...
between the respective TF binding domains and specific nucleotides of the TFBS from experimentally determined protein–DNA 3D structures. In fact, most of the pairs discovered are actually the binding cores from the TF binding domains and TFBS, respectively.

The proposed approach has great potential for discovering intuitive and interpretable rules of TF–TFBS binding mechanisms. Such rules are able to reveal TF binding domains, detailed interactions between amino acids and nucleotides, accurate TFBS sequence motifs, and help better understanding and deciphering of protein–DNA interactions. It also offers strategic help to reduce the labor and costs involved in wet-lab experiments. With increasing computational power and more sophisticated mining approaches, the proposed methodology can be further improved for discovering more intriguing TF–TFBS binding patterns and rules.

In the future, approximate associations will be considered to handle the experimental and biological noises, although the inevitable computational burden needs to be carefully handled, and much more efforts are needed to distinguish real signals from the large number of false positives introduced by loosening the pattern matching and clustering. Combinatorial associations between multiple TF and TFBS k-mers will also be another challenging topic. We will also seek further real applications of the approach on experimentally verifiable TF–TFBS bindings.

SUPPLEMENTARY DATA
Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS
The authors are grateful to the anonymous reviewers for their valuable comments. They would also like to thank Shaoke Lou, Kin-Cheung Ling and Leung-Yau Lo for their valuable help on preparing the study and the manuscript.

FUNDING
Research Grants Council of the Hong Kong SAR, China (project numbers: CUHK41407 and CUHK414708. Funding for open access charge: Block Grant Project of the Chinese University of Hong Kong, ref #2150591; Focused Investment Scheme D on Hong Kong Bioinformatics Centre, The Chinese University of Hong Kong (project number: 1904014).

Conflict of interest statement. None declared.

REFERENCES
1. Luscombe,N.M. and Thornton,J.M. (2002) Protein–DNA interactions: amino acid conservation and the effects of mutations on binding specificity. *J. Mol. Biol.*, 320, 991–1009.
2. Luscombe,N.M., Austin,S.E., Berman,H.M. and Thornton,J.M. (2000) An overview of the structures of protein-DNA complexes. *Genome Biol.*, 1, REVIEWS001.
3. Galas,D.J. and Schmitz,A. (1987) DNAse footprinting: a simple method for the detection of protein-DNA binding specificity. *Nucleic Acids Res.*, 5, 3157–3170.
4. Garner,M.M. and Revzin,A. (1981) A gel electrophoresis method for quantifying the binding of proteins to specific DNA regions: application to components of the Escherichia coli lactose operon regulatory system. *Nucleic Acids Res.*, 9, 3047–3060.
5. Smith,A.D., Sumazin,P., Das,D. and Zhang,M.Q. (2005) Mining ChIP-chip data for transcription factor and cofactor binding sites. *Bioinformatics*, 21(Suppl 1), i403–i412.
6. MacIsaac,K.D. and Fraenkel,E. (2006) Practical strategies for discovering regulatory DNA sequence motifs. *PLoS Comput. Biol.*, 2, c36.
7. Liu,X.S., Brutlag,D.L. and Liu,J.S. (2002) An algorithm for finding protein–DNA binding sites with applications to chromatin immunoprecipitation microarray experiments. *Nat. Biotechnol.*, 20, 835–839.
8. Matys,V., Kel-Margoulis,O.V., Fricke,E., Liebich,I., Land,S., Barre-Dirrie,A., Reuter,I., Chekmenev,D., Krull,M., Hornischer,K. et al. (2006) TRANSFAC and its module TRANScompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.*, 34, 108–110.
9. Hulo,N., Bairoch,A., Bulliard,V., Cerutti,L., Cuche,B.A., de Castro,E., Lachaize,C., Langendijk-Genevaux,P.S. and Sigrist,C.J.A. (2008) The 20 years of PROSITE. *Nucleic Acids Res.*, 36(Suppl 1), D245–D249.
10. Bateman,A., Coin,L., Durbin,R., Finn,R.D., Hollicy,V., GrifyRihes-Jones,S., Khanna,A., Marshall,M., Moxon,S., Sonhammer,E.L.L. et al. (2004) The Pfam protein families database. *Nucleic Acids Res.*, 32, D138–D141.
11. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weisig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, 28, 235–242.
12. Kel,A.E., Goessling,E., Reuter,I., Cheremushkin,E., Kel-Margoulis,O.V. and Wingender,E. (2003) MATCH: a tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res.*, 31, 3576–3579.
13. Stormo,G.D. (1988) Computer methods for analyzing sequence recognition of nucleic acids. *Annu. Rev. Biochem.*, 17, 241–263.
14. Jensen,S.T., Liu,X.S., Zhou,Q. and Liu,J.S. (2004) Computational discovery of gene regulatory binding motifs: a Bayesian perspective. *Statistical Science*, 19, 188–204.
15. Tompa,M., Li,N., Bailey,T.L., Church,G.M., Moor,B.D., Eskin,E., Favorov,A.V., Fritl,M.C., Fu,Y., Kent, et al. (2005) Assessing computational tools for the discovery of transcription factor binding sites. *Nat. Biotechnol.*, 23, 137–144.
16. Sandve,G.K., Abul,O., Walseng,V. and Drablos,F. (2007) Improved benchmarks for computational motif discovery. *BMC Bioinformatics*, 8, 193.
17. Jones,S., van Heyningen,P., Berman,H.M. and Thornton,J.M. (1999) Protein-DNA interactions: a structural analysis. *J. Mol. Biol.*, 287, 877–896.
18. Luscombe,N.M., Laskowski,R.A. and Thornton,J.M. (2001) Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. *Nucleic Acids Res.*, 29, 2860–2874.
19. Krishna,S.S., Majumdar,I. and Grishin,N.V. (2003) Structural classification of zinc fingers: survey and summary. *Nucleic Acids Res.*, 31, 532–550.
20. Jones,S., Shanahan,H.B., Berman,H.M. and Thornton,J.M. (2003) Using electrostatic potentials to predict DNA-binding sites on DNA-binding proteins. *Nucleic Acids Res.*, 31, 7189–7198.
21. Mandel-Gutfreund,Y., Schuler,O. and Margalith,H. (1993) Comprehensive analysis of hydrogen bonds in regulatory protein-DNA complexes: in search of common principles. *J. Mol. Biol.*, 253, 370–382.
22. Mandel-Gutfreund,Y. and Margalith,H. (1998) Quantitative parameters for amino acid-base interaction: implications for prediction of protein-DNA binding sites. *Nucleic Acids Res.*, 26, 2306–2312.
23. Sarai,A. and Kono,H. (2005) Protein-DNA recognition patterns and predictions. *Annu. Rev. Biophys. Biomol. Struct.*, 34, 379–398.
Appendix 1

Algorithm 1 Pseudocode of Apriori algorithm (29)

```
data: A dataset of itemsets
L_n: Frequent n-itemsets
C_n: Candidate n-itemsets
x: An itemset
mnsupport: Minimum Support

i ← 1;
Scan data to get L_i;
while L_i ≠ ∅ do
    \( L_{i+1} \leftarrow \text{EXTEND}(L_i); \)
    \( L_{i+1} \leftarrow \emptyset; \)
    For each \( x \in C_i \) do
        If \( \text{support}(x) \geq \text{mnsupport} \) then
            \( L_{i+1} \leftarrow L_{i+1} \cup \{x\}; \)
        end if
    end for
    i ← i + 1;
end while
```

Notes:

\( \text{EXTEND}(L) \) is the function ‘Candidate itemset generation procedure’ stated in (29).

\( \text{Support}(x) \) returns the support (30) of the itemset \( x \).

A frequent \( n \)-itemset is the \( n \)-itemset support is higher than \( \text{mnsupport} \).