2740–2746 *Nucleic Acids Research, 1998, Vol. 26, No. 11*

Computational approaches to identify leucine zippers

Erich Bornberg-Bauer*, Eric Rivals and Martin Vingron

Deutsches Krebsforschungszentrum, Theoretische Bioinformatik, Im Neuenheimer Feld 280, D-69120 Heidelberg, Germany

Received January 21, 1998; Revised and Accepted March 31, 1998

ABSTRACT

The leucine zipper is a dimerization domain occurring mostly in regulatory and thus in many oncogenic proteins. The leucine repeat in the sequence has been traditionally used for identification, however with poor reliability. The coiled coil structure of a leucine zipper is required for dimerization and can be predicted with reasonable accuracy by existing algorithms. We exploit this fact for identification of leucine zippers from sequence alone. We present a program, 2ZIP, which combines a standard coiled coil prediction algorithm with an approximate search for the characteristic leucine repeat. No further information from homologues is required for prediction. This approach improves significantly over existing methods, especially in that the coiled coil prediction turns out to be highly informative and avoids large numbers of false positives. Many problems in predicting zippers or assessing prediction results stem from wrong sequence annotations in the database.

INTRODUCTION

The identification of a particular transcription factor, the retrovirally transduced oncoprotein Jun (for a review see 1), and the characterization of its dimerization interface with a repeat of leucines by Landschulz *et al*. (2) in 1988 caused much interest in this motif. It was termed the 'leucine zipper' (LZ) and many sequences with such a leucine repeat were subsequently proposed to be LZs. Brendel and Karlin (3) argued that, since Leu is the most frequent amino acid, such a pattern may be easily found by chance. Consequently, many annotations in current databases may be wrong.

Leucine zippers are now commonly described as two-stranded, left-handed helical structures wrapped around each other (in a 'superhelix' or 'coiled coil'). They have a repetitive pattern where each leucine is followed by six other residues to form a heptad. Residues between the leucines may, in principle, be any amino acid. The number of heptads in general assumes a value from three to six, typically four. While coiled coils are known as long 2-, 3-, 4- or 5-stranded units in parallel or antiparallel configurations (4), natural LZs occur only as short parallel dimers. Since coiling reduces the repeat unit to 3.5, instead of 3.6 as observed in a normal α-helix, each two residues that are separated by six others in sequence would be superimposed in a helical wheel representation. In contrast to helical bundles, the fit between the two strands can be described by a 'knobs into holes' model (4). Residues in the leucine repeat are denoted *a*–*g*, starting with *d* for the first Leu. Positions *a* and *d* are mostly occupied by hydrophobic residues. They establish the helical interface at the core, which is viewed as the major stabilizing factor. Leu is important because of the flexible side chain. Positions *e* and *g* tend to be charged such that the ability to form salt bridges may significantly help to specify orientation and the dimerization partner. Consequently, specificity of dimerization depends on solvent conditions, such as pH. This is particularly meaningful for many oncoproteins, such as the competing pairs Jun–Jun and Jun–Fos (5) or Myc–Max and Max–Max (6). Positions *b*, *c* and *f* are solvent exposed and mostly occupied by rather hydrophilic residues, but poorly conserved.

LZs frequently occur together with DNA binding domains, e.g. in eukaryotic transcription factors. These proteins are involved in complex control circuits which govern gene expression during cell differentiation and tumour development. Proteins in which the leucine repeat occurs together with a so-called basic region (BR) as the DNA binding domain are known as basic zippers (or bZIP proteins). This class comprises well known proteins such as Fos, Jun, CREB/ATF, ATF2, 3 and 4 and AP1. Interestingly the co-occurrence of these motifs was found in a sequence alignment of Fos proteins alone and has been reported simultaneously with the LZ (7). A comprehensive review on bZIP sequences and their biological roles has recently been given by Hurst (1). In the BR only a few residues are strictly conserved. It is mostly assumed that LZs mediate dimerization. Then DNA binding takes place and transcription is started by allowing the RNA polymerase to bind both the transcription factor and the DNA. Binding to DNA is induced by two long α-helices, directly extending the LZ region and smoothly dividing to bind to the DNA (scissors grip model; 8). Other transcription factors, such as Mad, Max and Myc, dimerize with an abasic helix–loop–helix (bHLH) motif and bind to DNA by a basic region (9). This region, however, differs significantly from the BR in bZIP proteins. Their LZs often lack a strict leucine repeat. Some classes of protein kinases were reported to contain LZs (10). Since, however, no published experimental evidence for the coiled coil nature of these domains exists, we primarily refer to eukaryotic transcription factors in this work.

When sequences coding for the same structural motif share a common ancestor, a pattern of residues which is crucial for function and/or structure is likely to be conserved. Depending on

*To whom correspondence should be addressed. Tel: +49 6221 42 2725; Fax: +49 6221 42 2849; Email: bornberg@dkfz-heidelberg.de

the degree of conservation, various methods, such as pattern searches or alignment-based profile searches, can then be applied to determine family membership of a query sequence. Such methods will fail when an evolutionary relationship is either hardly detectable or even absent, e.g. as a result of convergent evolution. LZ-containing proteins appear in diverse families which lack a common evolutionary origin and sequence similarity among the LZs is extremely low. A pattern search in the PROSITE database using the repeated leucines performs worst of all patterns, producing hundreds of false positive and many false negative hits (11). Profile searches are very unreliable too (data not shown).

The program TRESPASSER (12) is reported to predict LZs with a high reliability. Starting with all SwissProt entries the authors first eliminated all non-LZ coiled coils from their data set. Two training sets were collected to derive patterns that are statistically indicative for zippers or non-zippers respectively. Assignment was based on the presence of a leucine repeat, a coil prediction, evidence of dimer formation and DNA/RNA binding ability. A coil prediction program (13) was used to align all sequences. The authors report 18 false negatives and 36 false positives of COILS in predicting annotated LZs with a leucine repeat as a coiled coil. From both data sets tuple, frequencies were collected and re-filtered manually. To predict a query sequence as zipper or non-zipper, occurrences in both sets are summed to derive a score. Presence of a leucine repeat (three heptads or more) seems to be mandatory for positive identification. When tested on this particular set of 'potential zippers', TRESPASSER is reported to show a trade-off between false positives (down to 0%) and false negatives (down to 3%) which can be tuned by the ratio of pairs from both sets. This approach, however, leaves considerable room for improvement, since the underlying database annotations are not always correct and the leucine repeat need not be strictly conserved.

The problems with pattern- or profile-based approaches led us to exploit the coiled coil structure of a LZ for prediction purposes. We show that such a method combining coil prediction with the search for a leucine repeat works almost perfectly for bZIP and bHLH-LZ proteins, and that a leucine repeat pattern that is too strict often fails to find known LZs. The Achilles' heel for any such prediction method seems to be judgement by database annotations.

MATERIALS AND METHODS

Sequences were retrieved from SwissProt (14) release 34, having 59 021 entries. Based on the annotations we distinguished the following three classes of sequences and, where in doubt, consulted the literature: (i) annotated zippers, denoting the 152 entries that are annotated as 'leucine zipper'; (ii) undecided, comprising 41 entries annotated as 'leucine zipper like', '... by similarity', etc., which instances were obviously classified by exactly the criteria we investigate; (iii) annotated non-zippers, referring to all entries whose annotation makes no mention of a LZ.

Leucine repeat pattern

As pointed out in the Introduction, LZs are commonly identified by their leucine residues at a fixed spacing of seven residues. Comparing sequences in annotated and undecided zippers, one finds that one leucine in the repeat has frequently been replaced, mostly by Met, Val or Ile. This led us to distinguish a class of strict and one of relaxed occurrence of the leucine repeat. We call a strict leucine repeat one where there are at least five leucine residues with the prescribed spacing, i.e. four repeats. A relaxed leucine repeat is one where any of the five positions is mutated either to Met, Val or Ile. In fact, there are strict leucine repeats which are even longer. Some relaxed leucine repeats extend over five repeats with the one mutated position in the middle. Overall, 4318 leucine repeat patterns, either strict or relaxed, were found in 3370 different sequences. Sometimes repeats overlap or a sequence contains more than one repeat.

Other domains

It can be helpful to study the DNA binding or other protein–protein interaction domains residing next to a predicted LZ. Patterns describing such domains are taken from Prosite (14). In this context we use the following abbreviations: BR stands for a basic region of the bZIP-type proteins (102 instances in SwissProt 34), a pattern which is between 14 and 16 residues long (Prosite entry PD00036); bHLH denotes a helix–loop–helix motif (222 instances), which is characterized by a pattern derived from the second helix (PD00038).

Coil prediction

Two well-known methods are available for coiled coil prediction from sequence information. They are more general in scope because coiled coils appear in many different proteins, such as myosins, intermediary filaments, keratin, CAP, tRNA synthetase, G proteins, kinases, etc. (4,13,15). COILS, described by Lupas *et al*. (13) in 1991, and its subsequent version COILS2 are profile searches based on work by David Perry (16). Although their parameters were derived from fibrous proteins they work fairly well for most other coiled coils (17). A newer method that works well, especially for long two-stranded parallel coils, is the program PAIRCOIL by Berger *et al*. (1995). It is based on the correlated occurrence or non-occurrence of residues throughout the heptads.

We use both available programs for coil prediction. COILS is preferred for several reasons. First, according to Lupas (15) COILS tends to slightly over-predict, which is good for combination with another criterion, while PAIRCOIL does not perform so well, especially for short coiled coils. Next, the complete source code is available such that the program could be better adapted to our needs. Also, it is readily implemented in numerous database search programs. We collect frames, i.e. start and end positions of coil predictions and their maximum probabilities, stepped in units of 0.1 from 0.0 to 1.0. Sequences with a probability $P_c > 0.5$ are considered as ones with a coiled coil. Calculations for the positional probability are obtained using the newer version COILS2, which gives comparable results but slightly lower predictions for short coils in general. PAIRCOILS (18) was used for comparison with all parameters set to default. Predictions agreed fairly well with the results from COILS, where the latter yields slightly more and longer predicted frames.

Definition of a LZ and the recognition algorithm (2ZIP)

As mentioned above, a LZ is not simply a coiled coil but has several special features. First, the helices in a LZ form a parallel dimer and this orientation is determined by the leucines (for a

review see 4). Also, LZs comprise only four to six heptads, which is the minimum number required for a stable dimer (19). This is essential since, in contrast to coiled coils in fibrous proteins, the dimer must not be too stable, to allow for reversible dimerization and a flexible choice of dimerization partners. This readily leads to a working definition of a LZ as a short, parallel, dimeric coiled coil, generally containing five to seven leucines with a characteristic repeat of length seven. The precise number of leucine residues is flexible in as far as it seems to govern stability and orientation of dimerization. This definition can be directly translated into a simple algorithm, which is assessed in this paper.

First the coiled coil prediction is computed for a given sequence. Assume a sequence contains either a strict or a relaxed leucine repeat. We demand a minimal overlap of 21 residues between the region predicted as coiled coil and the leucine repeat. We need to exclude long structural coiled coil proteins that happen to have a few leucines with the right spacing. The algorithm rejects sequences where the coiled coil prediction stretches over >90 residues or where there are more than three coiled coil regions predicted. Of course, these instances may well represent a LZ in a strictly biophysical sense, but they do not meet the functional criteria that 'define' a LZ. Depending on whether the basis of the prediction is a strict or a relaxed leucine repeat, this scheme will produce two classes of predicted LZs.

Availability

Programs (which invoke COILS), results (in well-formatted files) and an on-line server can be found on the web at: http://www.dkfz-heidelberg.de/tbi/services/2zip/2zip

RESULTS

Predicted and annotated zippers

We searched for sequences matching the mentioned criteria for leucine repeats and coil prediction. We recorded all frames of coil predictions for all annotated and undecided LZs as well as all other sequences containing a leucine repeat. First we inspected all frames with both a leucine repeat and a coiled coil prediction. In the following, sequences containing at least one zipper according to the above-mentioned criteria are termed predicted zippers, while all others are called predicted non-zippers. To keep figures consistent we decided in the statistics below only to count each such sequence once, even if more than one frame with a prediction was found.

In an initial assessment we referred to the database annotations to verify our results. Table 1 summarizes the distribution of annotations among the predicted zippers and non-zippers. Overall 3398 sequences contain a leucine repeat frame, a zipper annotation or both. From this base set, 408 leucine repeat frames overlap sufficiently with a coil frame to be predicted as LZs. 121 of them are annotated as zippers. Nearly half of these require the relaxed leucine repeats to be detected. On the other hand 276 (247 + 29) sequences whose annotations make no mention of a LZ are predicted to contain a LZ. Based on annotation these would be the false positives when one combines predictions based on relaxed and strict leucine repeats. Most of these false positives (247) stem from admitting relaxed leucine repeats. Below we are going to discuss the 29 'false positives' with a strict pattern in detail. A summary is given in Table 2. At the top of the list is a member of the Myc family also having a bHLH motif (MYC_AVIMC). This makes it highly likely that our prediction is in fact correct. The next one is a transcription factor and possesses a BR domain (YDC3_SCHPO). Since coupling of a LZ and a BR domain is frequently observed it appears that this protein too might indeed contain a LZ. Two more sequences are nuclear and annotated as transcription factors, which again makes the LZ prediction a feasible guess (STA4_MOUSE and NEK2_HUMAN). The other sequences, 17 eukaryotic and 8 prokaryotic proteins, may well be real false positives since, to our knowledge, there is no reason to assume that any of them might have a LZ. It is noteworthy that only two of the false negatives are short fibrous proteins missed by the filtering procedure. Some sequences that would have been very tempting to predict as containing a LZ are not in our false positive list. An example is DNA topoisomerase 2. It binds DNA and has both a leucine repeat and a coiled coil, but without overlap of the two corresponding frames (15). Incidentally, this example was used by Hirst *et al.* (12) in the positive list for deriving rules to predict LZs. In our data set all four occurrences of DNA topoisomerase 2 are classified as non-zippers.

Table 1. Predicted and non-predicted zippers and their annotations (see text for explanation)

Prediction	Annotation	Relaxed	Strict	No repeat	All
Zipper	Annotated zippers	54	67		121
	Undecided	7	4		11
	Annotated non-zippers	247	29		276
	Sum	308	100		408
Non-zipper	Annotated zippers	7	3	21	31
	Undecided	7	2	21	30
	Annotated non-zippers	2543	358		2901
	Sum	2557	363	42	2962

For nearly 3000 sequences our method does not predict a LZ. Most sequences are rejected by our method, in accordance with the annotation. For these sequences consideration of the coiled coil prediction really helps in rejecting a zipper hypothesis which is based on a leucine repeat alone. There are 31 sequences where prediction failed. Table 3 gives a compilation of these instances that are, according to the annotations, false negatives. Interestingly, all of them lack a coil prediction. Only three had a strict repeat pattern; eight had a relaxed one. All others had a repeat that was even more mutated, mostly with the variable residues from the relaxed pattern, Met, Ile and Val. Eight have a basic region which suggests that they are bZIP proteins and were missed by our procedure. On the other hand, in seven instances the BR does not have the right spacing such that coincidence or other functions cannot be excluded. For three of the BR-containing sequences that were annotated as containing a LZ, this zipper was shorter than four heptads, which is commonly viewed as the minimum number for a stable dimer. Four had two of the five leucines replaced by other residues and two had one substitution with non-canonical residues (Ala or Tyr). None had a coil prediction by COILS and only four by PAIRCOILS.

Table 2. Predicted but not annotated leucine zippers with a strict repeat pattern

ID	H	DOM	FRAME	SEQ	COMM
MYC_AVIMC	4	bHLH	389-417, 422;	LKKATEYVLSLQSDEHKLIAEKEQLRRRREQLKHNL	nucl., binds DNA, myc-homology, ETF
YDC3_SCHPO	4	BR	292-320, 330;	LERTAKELTEKVAILETRVRELEMENNWL	nucl., DNA-binding, ETF
STA4_MOUSE	5		252-280, 749;	LHNGLDQLQNCFTLLAESLFQLRQQLEKL	nucl., signaltransduction, ETF
NEK2_HUMAN	4		306-334, 445;	LKLKEIQLQERERALKAREERLEQKEQEL	nucl., S/T-kinase, cell regulation
APAR_PIG	4		160-188, 202;	LMQCLPNLEEIKLALELYKLSLETKLLEL	extracellular, apolipoproteins
ASE1_YEAS	5		495-530, 885;	LRNSATTLQEDELLLETCENELKRLEEKLTLYKPIL	in mitosis
CD72_HUMAN	4		86-114, 359;	LPCRTTCLRYLLLGLLLTCLLLGVTAICL	cell., S-S linked homodimer
CD72_MOUSE	4		143-171, 354;	LREKISQLGQKEVELQESQKELISSQDTL	cell., S-S linked homodimer
DPO1_THETH	4		458-486, 834;	LQALSLELAEEIRRLEEEVFRLAGHPFNL	prok., DNA - polymerase
EF1D_ARTSA	5		58- 93, 237;	LSNKVEALSSENKELKKCIDGLQGLLLGLRQRIETL	nuclear, elongation factor
EF1D_HUMAN	5		80-115, 281;	LVVRIASLEVENQSLRGVVQELQQAISKLEARLNVL	nuclear, elongation factor
EF1D_RABIT	5		80-115, 280;	LAVRIASLEVENQNLRGVVQDLQRAVSKLEARLSAL	nucl., elongation factor
EF1D_XENLA	5		58- 93, 265;	LAARVANLEQENQSLHKVVKDLQSAISKLESRLSTL	nucl., elongation factor
GVPJ_APHFL	4		79-107, 248;	LSTKAQRLVEENQQLQHRLESLEAKLNSL	prok.
HLYD_PASHA	5		253-288, 478;	LLAQENKLIEAQNELAVYRSKLNELENDLLNVKEE	prok.
IL11_MOUSE	4		119-147, 199;	LKTLEPELGALQARLERLLRRLQLLMSR	cytokine, low prob., no PAIRCOIL-frame
K1C3_XENLA	4		192-220, 280;	LRRVLDELTLARGDLEMQIESLTEELAYL	fibrous protein (IF, keratin)
LECH_RAT	5		107-142, 283;	LEKHQEDLREDHSRLLLHVKQLVSDVRSLSCQMAAL	memb., mediates endocytosis
MMGL_MOUSE	4		135-163, 304;	LKTDLSDLTDHVQQLRKDLKALTCQLANL	memb., involved in tumorgenesis, ETF
OMPA_THEMA	5		79-114, 400;	LAGASGDLAQVVGNLSDKYMALEEKVNGLTGILDTL	memb., fibrous
PAC1_YEAST	4		95-123, 494;	LQKKIIELEQNTETLVSQIKDLNTQVSEL	transducin
PDP_BACSU	4		269-297, 434;	LHELVLTLGSQMVVLAKKADTLDEARAKL	prok., phosphorylase
PRE_BACSP	5		300-335, 415;	LKKEVKELRSTNKSLSEENGRLKSTVEHLTNEIESL	prok., DNA-binding
RPOD_SORBI	4		272-300, 435;	LEDEYRTLE DEYETLEDEY GILEDEYRTL	prok., chloroplast, RNA-polase subunit
SUM1_YEAST	5		114-142, 1062;	LLSKDTSLTDSVQDLFNSLKVLSHNQSVL	—
SYA_HAEIN	5	ZnF	706-734, 874;	LHNQQRILTQSADLLKSDVNTLAEKIQQL	prok., homotetramer, A-tRNA-synthetase
VIRB_SHIFL	4	HTH	196-224, 309;	LFNYYKGLEKANESLSSTLPILKEEIKDL	prok., DNA-binding, trc. regulation
YA65_MOUSE	4		296-324, 472;	LQMEKERLRLKQQELFRQELALRSQLPTL	binds to yes-kinase
YWFN_BACSU	4		180-208, 258;	LKMENERLKKENQELQNKTEQLEAEVQKL	prok.

Explanation of abbreviations: ID, SwissProt ID number; H, whether a 4-heptad or 5-heptad was present in the annotated frame; DOM, whether another domain, indicative of a eukaryotic transcription factor is present (BR, basic region; HLH, helix–loop–helix; HTH, helix–turn–helix; ZnF, zinc finger); Frame, start and end position of LZ frame, overall length of the protein; SEQ, sequence in the frame; COMM, the comments in the database, as far they were helpful: cellular location (nucl., nuclear; memb., membrane; prok., prokaryotic), molecular function and cellular function (trc., transcription) as well as additional information (ETF, eukaryotic transcription factor).

Another five belong to the Myb class of eukaryotic transcription factors, where two of the five leucines are substituted. Serious concerns must be raised about the zipper nature of their annotated regions. The proposed zipper region is far from the DNA binding domain and thus does not resemble the known architecture of other zipper-containing transcription factors. Finally, even more strikingly, it was experimentally shown that under physiological conditions the domain does not form an α-helix, not to mention a dimer (20).

Interestingly, for eight of the 31 sequences the corresponding references do not mention a LZ. For some putative instances it appears unlikely they are LZs because of their biological function or their sequences, or both. Some have strong helix breakers (Gly or Pro) in the proposed α-helical regions. Others are either membrane proteins or their leucine repeat is at the very N-terminus, which is never observed for generally accepted LZs. Some are definitely annotated as not binding to DNA. This of course does not exclude the existence of a LZ.

Association with other domains

Due to the problems with database annotations, we searched for other criteria to recognize whether a sequence might contain a LZ. As mentioned above, LZs frequently occur together with a DNA binding basic region or a HLH domain. Therefore, we combined our criteria with the use of regular expressions to search for these regions adjacent to or in the N-terminus of transcription factors. When order and spacings of the motifs (i.e. BR or bHLH + predicted zipper) are correct we can be sure that these instances are indeed zippers. We refer to such zippers that co-occur with a basic region or a bHLH domain as confirmed zippers. LZs may also be associated with homeodomains in some homeobox families (21). Since, however, these cases seem to be restricted to plants (e.g. *Arabidopsis thalliana*, hat1-thal, hat2-…) and the structural role of the associated LZ is somewhat unclear, we do not refer to such motifs.

Table 4 shows the distribution of sequences with a co-occurrence of both domains among the predicted zippers and predicted non-zippers. With one exception, all predictions of LZs in confirmed zippers were correct when the BR motif was used for evaluation. Slightly less reliable results are achieved when the bHLH motif is used, such that we find an overall accuracy of at most 3% false negatives. These results constitute further confirmation of our prediction method and strengthen the view that combining a relaxed leucine repeat and a coiled coil prediction are a good strategy to identify LZs.

Table 3. Predicted non-zippers which are annotated

ID	P	DOM	ETF	FRAME	PATT	SEQ	COMMENTS
ATF6_HUMAN	n	BR	ATF	33 - 61, 68;	LALLV	LEARLKAALSENEQLKKENGRLKRQLDEV	nucl., binds DNA(CRE) as a dimer
BBF2_DROME	n	BR		468 - 503, 515;	LLYLLL	LERRVEILVTENHDYKKRLEGLEETNANLLSQLHKL	nucl., trc. activator, binds enhancers
CAD1_YEAST	n	BR		71 - 99, 409;	LLNLL	LQERVELLEQKDAQNKTTTDFLLCSLKSL	nucl. (?)
CBL_HUMAN	n			857 - 892, 906;	LLYLIL	LSSEIENLMSQGYSYQDIQKALVIAQNNIEMAKNIL	dito
CBL_MOUSE	n			847 - 882, 896;	LLYLIL	LSSEIERLMSQGYSYQDIQKALVIAHNNIEMAKNIL	nucl.
CNC_DROME	n	BR		387 - 408, 533;	LLIL	LNQDRDHLESERKRISNKFAML	morphogenesis control, DB
CPR2_PETCR	n	BR		218 - 253, 393;	LLLIAL	LETQVSQLRVENSSLLKRLTDISQRYNDAAVDNRVL	nucl., binds DNA as a dimer
CPR3_PETCR	n	BR		224 - 259, 296;	LLLIVL	LQERLDNLSKENRILRKNLQRISEACAEVTSENHSI	dito
CSE2_YEAST	n	(BR)		126 - 147, 149;	ILLL	IHQREQELQIKRDVLDDLYRKL	nucl.
GCF_HUMAN	n			359 - 380, 784;	LMLL	LKQAMTFMKRRQDELKHESTYL	nucl., DNA binding, represses trc., DB
GIAN_DROME	n	BR		412 - 433, 448;	ILLL	IAIRAAYLERQNIELLCQIDAL	nucl., gene regulation in morphogenesis
IN35_HUMAN	s			5 - 26, 282;	LLLLL	LDAALHALQEEQARLKMRLWDL	
MT28_YEAST	r			140 - 161, 187;	LLLW	LNTQINKLRDRIEQLNKENEFW	nucl.(?), binds promoter, no trc. activation
MYB_HUMAN	n		MYB	376 - 397, 640;	MILL	MIVHQGTILDNVKNLLEFAETL	nucl., DNA-binding, trc. activator
MYB_MOUSE	n		MYB	375 - 396, 636;	MILL	MIVHQGTILDNVKNLLEFAETL	dito
MYB_AVIMB	n		MYB	305 - 326, 382;	MILL	MIVHQSNILDNVKNLLEFAETL	dito , DB
MYB_CHICK	n		MYB	376 - 397, 641;	MILL	MIVHQSNILDNVKNLLEFAETL	dito , DB
MYB_BOVIN	n		MYB	375 - 396, 640;	MILL	MIFHQSTILDNVKNLLEFAETL	dito , DB
NAPT_HUMAN	s			348 - 369, 639;	LLLLL	LAVGLILLAGSLVLLCTCLILL	memb., L-rich, 2 Gly, 20 annotations, DB
NAPT_RAT	s			346 - 367, 637;	LLLLL	LAVGLILLAGSLVVLCTCLILL	dito , DB
OCT2_HUMAN	r			388 - 409, 478;	LLLL	LSQASSSLSTTVTTLSSAVGTL	TF in B-cells
OPI1_YEAST	r			139 - 60, 404;	LLLL	LVTCLHLLKLANKQLSDKISCL	anabolic gene regulation
PCR1_SCHPO	n	BR		42 - 66, 171;	ASLLL	ANAAFEQSKRLQLLLSQLQQEAFRL	nucl. (?), regulatory
RELB_MOUSE	n			22 - 50, 558;	LVLIL	LSSLSLTVSRTTDELEIIDEYIKENGFGL	nucl., stimulates promoter, not DNA binding
RELB_HUMAN	n			40 - 68, 579;	LVLIL	LSSLSLAVSRSTDELEIIDEYIKENGFGL	dito
RNBP_RAT	n			185 - 206, 419;	LLML	LLNLVEQLGEEDEEMTDKYAEL	renin binding, homo-dimer
RNBP_HUMAN	r			185 - 206, 417;	LLLL	LLNLVEQLGEADEELAGKYAEL	renin binding, forms homo-dimers
RPC3_YEAST	r			581 - 602, 654;	LLLL	LEWNMANLLFKKEKLKQENSTL	nucl., RNA-polymerase III subunit
SRS2_YEAST	r			222 - 243, 1175;	LLLL	LLMYTFRLLTRVRVLSNIKHVL	ATP-dependent DNA helicase SRS2, DB
TGAB_TOBAC	n			211 - 232, 242;	LMLI	LEDKVRIMHSTIQDLNAKVAYI	nucl., DNA binding
VP2_ROTBR	r			665 - 686, 880;	LLLL	LRDRLRLLPVEVRRLDIFNLIL	core of bovine rotavirus, binds NAs, DB

Explanation of abbreviations (where different from Table 2): P, pattern in the annotated frame; r, relaxed; s, strict; n, none; ETF, family of eukaryotic transcription factor as which the sequence is annotated; PATT, leucine repeat pattern; DB, the corresponding LZ annotation cannot be found in the original literature.

Table 4. Confirmed and predicted zippers: occurrences of sequences with additional domains, indicative of eukaryotic transcription factors (bZIP, bHLH-LZ) with the correct spacing (another eight annotated zippers had a BR but not with the correct spacing)

Prediction	Annotation	bZIP	bHLH-LZ	Other
Zipper	Annotated zippers	57	17	44
	Undecided	4	1	6
	Annotated non-zippers	1	1	277
	Sum	62	19	327
Non-zipper	Annotated zippers	0	0	31
	Undecided	0	5	25
	Annotated non-zippers	0	1	2928
	Sum	0	6	2984

Strength of coil prediction in confirmed LZs

In the following, we evaluate the positional coil probability for some of the confirmed LZs. This has important structural implications because it was reported that the basic region of bZIP proteins shows a slight coiled coil probability (4). Likewise, the bHLH region of bHLH-LZ proteins has two α-helices adjacent to the LZ region. This suggests that at least the hydrophobic interface would be similar in a bundle and a zipper and the border between the domains difficult to detect. Finally, several zipper motifs were reported to be relatively unstable. For the case of Myc–Myc dimers, neither the complete dimer nor the zipper fragments alone are stable and also the dimer of the Myc–Max LZs alone is less stable than the LZ dimer of bZIP proteins alone (19). This appears reasonable when one considers that bHLH-LZ proteins have two motifs to specify dimerization, the bHLH and the LZ domain. Most probably the LZ motif serves primarily for recognition and not so much as a stabilizing domain (9). In Figure 1 we report the coil prediction strength at each position separately for the members of the bZIP class families Jun, Fos and ATF, as well as for Myc from the bHLH-LZ class. ATF and Fos show the expected behaviour with 100% coil probability in the zipper frame and a sharp drop-off at the border of the adjacent regions. There are, however, significant deviations for Jun and Myc. The BR of Jun shows a very high coil probability. This is interesting considering that Jun, unlike Fos, can form homodimers and thus the coil region may be extended

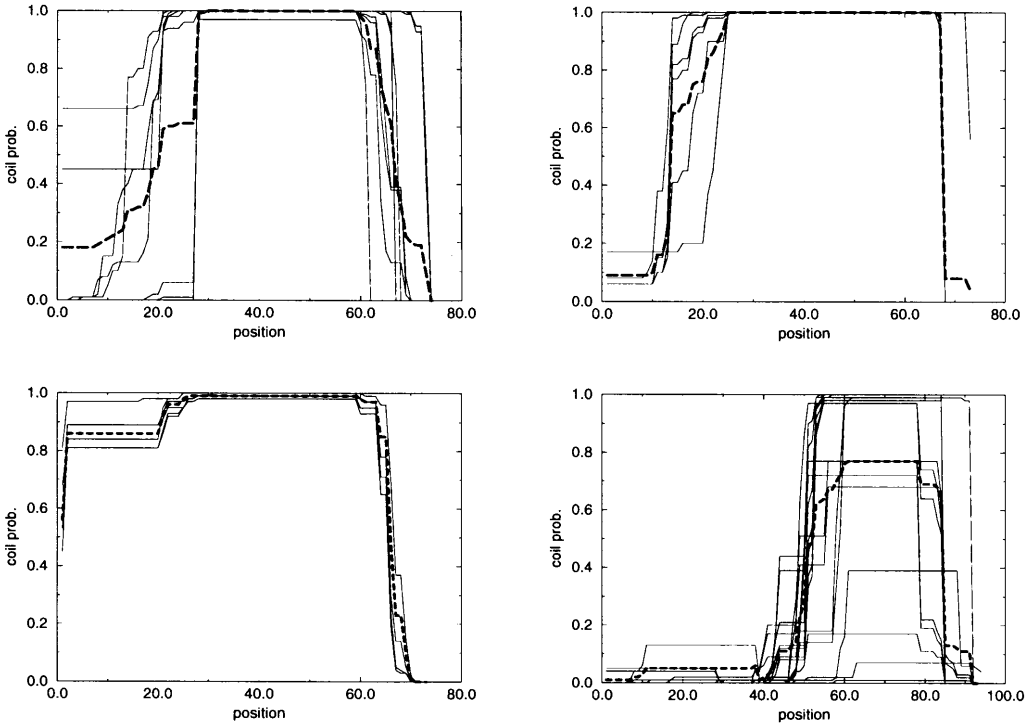

Figure 1. Coil probabilities at each position of bZIP and bHLH-Z proteins calculated with COILS2. All frames are aligned such that the leucine repeats within a family overlap. The leucine repeat starts at position 30 for the bZIP proteins ATF (top left), Fos (top right) and Jun (bottom left) and at position 60 for the bHLH family Myc (bottom right). Averages are given as thick dashed lines.

in the absence of DNA. This complies well with the fact that the BR was shown to 'zip up' when an adequate dimerization partner with an extended LZ region is provided (22). Further, it can be clearly seen that the LZs of Jun and Fos are in fact six heptads long (instead of four as is mostly annotated). In both cases the motif is characterized by a His instead of the sixth Leu. With Fos the drop-off sharpens because of a Pro at the very end of the six heptads. Myc sequences indeed show widely varying coil probabilities. This may be due to a smaller contribution to dimer formation and strengthens the view that the LZ in bHLH-LZ proteins mediates specificity rather than stability.

DISCUSSION

We have presented a simple computational approach for identification of LZs by combining a standard coiled coil prediction algorithm with an approximate search for the characteristic leucine repeat. Another goal of our study was a systematic investigation of the co-occurrence of leucine repeats and a detectable coiled coil in LZs, in particular for eukaryotic transcription factors. To avoid the pitfalls of wrong annotations we use additional biological signals, such as DNA binding motifs, for verification. In summary, we find the following conclusions of both practical relevance and general interest.

First of all we designed a fast and easily applicable strategy to predict LZs. Specifically, for eukaryotic transcription factors the method has excellent accuracy. 2ZIP should be particularly useful to obtain a first guess about the presence of a LZ in a given sequence. Also, for sequences with no or few known homologues it will prove useful to decide whether the protein may dimerize through a LZ or not.

Secondly, in order to describe all annotated zippers in SwissProt it appears to be important to use different, 'relaxed' patterns, which of course increases the number of 'false positives'. Hirst *et al.* (12) circumvent this problem by confining their search to sequences with a strict leucine repeat. We show that such a criterion is a major source of false negatives.

Finally, it is interesting to observe that the biological needs for flexibility or alternative dimerization are well reflected by different coil probabilities. This is concluded from the coil probabilities of the basic region in bZIP proteins and the reduced coil probabilities of bHLH-LZ proteins. This is an intriguing observation because there are no thermodynamic or other biophysical considerations directly included in the coil prediction heuristics. Thus a 'false negative' from a coil prediction need not be a failure of the program, but may well reflect biophysical functionality.

Our work also pinpoints some basic problems in the fields of structure prediction and motif recognition. The need for flexibility in biological activity frequently results in marginal stability, which in turn leads to fuzzy rules of recognition. This undermines efforts at clear and precise definitions and sharp discrimination, e.g. between coiled coil and non-coiled coil or between LZ and coiled coil. Because 'LZs' with a three heptad and three of four leucine positions substituted by other residues have been postulated (1), classification of a LZ becomes basically a question of definition. Thus one may also conclude that there is no specific code for a LZ, but LZs are simply short parallel dimeric coiled

coils with as much additional regularity as needed for proper function, such as orientation and flexibility. We conclude that since a coiled coil prediction seems to be a more reliable indicator for a LZ, the hallmark of a LZ is rather the coiled coil than the leucine repeat. A prediction strategy making use of both features has a surprisingly high success rate, yet an ultimate classification of such proteins can only be achieved by homology comparison and structural information.

ACKNOWLEDGEMENT

Stimulating discussions with Andrei Lupas are gratefully acknowledged.

REFERENCES

1 Hurst,H. (1994) *Protein Profile*, **2**, 105–106.
2 Landschulz,W.H., Johnson,P.F. and McKnight,S.L. (1988) *Science*, **240**, 1759–1764.
3 Brendel,V. and Karlin,S. (1989) *Nature*, **341**, 574–575.
4 Lupas,A. (1996) *Trends Biochem. Sci*., **21**, 375–382.
5 O'Shea,E.K., Rutkowski,R. and Kim,P.S. (1992) *Cell*, **68**, 699–708.
6 Muhle-Goll,C., Nilges,M. and Pastore,A. (1995) *Biochemistry*, **34**, 13554–13564.
7 Vingron,M., Nordheim,A. and Muller,R. (1988) *Oncogene Res*., **3**, 1–7.
8 Brandon,J. and Tooze,C. (1991) *Introduction to Protein Structure*. Garland Publishing, New York, NY.
9 Littlewood,T.D. and Evan,G.I. (1994) *Protein Profile*, **1**, 639–709.
10 Dorow,D.S., Devereux,L., Dietzsch,E. and DeKretser,T. (1993) *Eur. J. Biochem*., **15**, 701–710.
11 Bairoch,A., Bucher,P. and Hofmann,K. (1993) *Nucleic Acids Res*., **25**, 217–221.
12 Hirst,J., Vieth,M., Skolnick,J. and Brooks,C. (1996) *Protein Engng*, **9**, 657–662.
13 Lupas,A., vanDyke,M. and Stock,J. (1991) *Science*, **252**, 1162–1164.
14 Bairoch,A. and Apweiler,R. (1992) *Nucleic Acids Res*., **25**, 31–36.
15 Lupas,A. (1997) *Curr. Opin. Struct. Biol*., **7**, 388–393.
16 Perry,D.A. (1982) *Biosci. Rep*., **2**, 1017–1024.
17 Lupas,A. (1996) *Coil 2.2 Manual*, at ftp://FTP.BIOCHEM.MPG.DE/Coils
18 Berger,B., Wilson,D.B., Wolf,E., Tonchev,T., Milla,M. and Kim,P.S. (1995) *Proc. Natl. Acad. Sci. USA*, **92**, 8259–8263.
19 Muhle-Goll,C., Gibson,T., Schuck,P., Schubert,D., Nalis,D., Nilges,M. and Pastore,A. (1994) *Biochemistry*, **33**, 11296–11306.
20 Ebneth,A., Adermann,K. and Wolfes,H. (1994) *FEBS Lett*., **337**, 265–268.
21 Rubert,I., Sessa,G., Lucchetti,S. and Morelli,G. (1991) *EMBO J*., **10**, 1787–1891.
22 Krylov,D., Olive,M. and Vinson,C. (1995) *EMBO J*., **14**, 5329–5337.