frontiers in Bioengineering and Biotechnology

REVIEW
published: 24 September 2015
doi: 10.3389/fbioe.2015.00143

Tandem repeats in proteins: prediction algorithms and biological role

*Marco Pellegrini**

Laboratory for Integrative Systems Medicine (LISM), Istituto di Informatica e Telematica, and Istituto di Fisiologia Clinica, Consiglio Nazionale delle Ricerche, Pisa, Italy

Tandem repetitions in protein sequence and structure is a fascinating subject of research which has been a focus of study since the late 1990s. In this survey, we give an overview on the multi-faceted aspects of research on protein tandem repeats (PTR for short), including prediction algorithms, databases, early classification efforts, mechanisms of PTR formation and evolution, and synthetic PTR design. We also touch on the rather open issue of the relationship between PTR and flexibility (or disorder) in proteins. Detection of PTR either from protein sequence or structure data is challenging due to inherent high (biological) signal-to-noise ratio that is a key feature of this problem. As early *in silico* analytic tools have been key enablers for starting this field of study, we expect that current and future algorithmic and statistical breakthroughs will have a high impact on the investigations of the biological role of PTR.

Keywords: proteins, tandem repeats, biological significance, protein TR detection algorithms, protein TR properties

OPEN ACCESS

Edited by:
John Hancock,
The Genome Analysis Centre, UK

Reviewed by:
Silvio C. E. Tosatto,
University of Padua, Italy
Michelle M. Simon,
Medical Research Council, UK

***Correspondence:**
Marco Pellegrini,
Istituto di Informatica e Telematica,
Consiglio Nazionale delle Ricerche,
Via Moruzzi 1, Pisa 56124, Italy
marco.pellegrini@iit.cnr.it

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: *12 June 2015*
Accepted: *07 September 2015*
Published: *24 September 2015*

Citation:
Pellegrini M (2015) Tandem repeats in proteins: prediction algorithms and biological role. Front. Bioeng. Biotechnol. 3:143. doi: 10.3389/fbioe.2015.00143

Introduction

A seminal paper (Andrade et al., 2001) reports the observation that repetitive subsequences that appear in tandem repetitions (TR) within the protein primary sequence often form integrated assemblies when these residues are mapped to their corresponding three-dimensional folded conformation. These TR confer multiple binding opportunities and may play a structural role by giving rigidity to a protein, and by exposing functional domains. Moreover, Andrade et al. (2001) remark that *tandem repeated structures* should not be assimilated to the traditional notions of *domains* and *motifs* that may appear singly or in multiple interspersed copies in each protein (while they can be repeated across families of protein), since they constitute a rather distinct class. They also remark that repeats in protein sequences are usually hard to detect because on average the repeating unit is relatively short, and moreover there can be considerable sequence divergence among units of the same TR. We will refer throughout this article to these repetitive sub-sequences as *Protein Tandem Repeats* (PTR or Protein-TR, for short).

A study by Marcotte et al. (1998) indicates that internal subsequence repetitions in protein primary structure are quite widespread. They have been detected in about 14% of all the then known proteins, with eukaryotic proteins being three times more as likely to have internal repeats than prokaryotic ones. More recent measurements in (Pellegrini et al., 2012) give a count of about 25% of the proteins in the Uniprot database (Apweiler et al., 2004) holding a PTR of length at least 20 aa.

A recent survey of some algorithmic aspects of PTR detection in protein sequences is in Luo and Nijveen (2014). In this survey, we will touch lightly on the multi-faceted aspects of PTR

research, including prediction algorithms, databases, early classification efforts, mechanisms of PTR formation and evolution, and synthetic PTR design. We also touch on the rather open issue of the relationship between PTR and flexibility (or disorder) in proteins.

Protein-TR Detection Algorithms Based on Sequence

Structural and functional properties of Protein-TR are often preserved also in presence of high divergence among the subsequences corresponding to the PTR units, both at the level of DNA coding sequence and at the level of AA sequence. This property makes automatic PTR detection a challenging task, and a variety of approaches have been implemented since the late 1990s. More recently, a tendency to integrating basic sequence data with evolutionary or biochemical annotations has emerged. **Table 1** reports the list of sequence-based algorithms.

Interestingly, early algorithms by Marcotte et al. (1998), Pellegrini et al. (1999), and Andrade et al. (2000) were instrumental to the first PTR classification efforts, while more recent tools have been aimed at providing web-server-based utilities, or at populating databases.

REP in Andrade et al. (2000) is one of the first PTR detection algorithms which uses a homology-based method to identify statistically significant protein repeats.

Other early methods developed for finding TRs in proteins are based on detecting sub-optimal alignments in the self-alignment matrix generated by the Smith-Waterman algorithm (or similar methods). Some methods developed along this line are `Internal Repeat Finder` (Marcotte et al., 1998; Pellegrini et al., 1999), `prospero` (Mott, 1999), RADAR (Heger and Holm, 2000), REPRO (Heringa and Argos, 1993; George and Heringa, 2000), and TRUST (Szklarczyk and Heringa, 2004). These methods often detect both tandem and interspersed repeats.

XSTREAM (Newman and Cooper, 2007) uses a seed expansion approach, while Jorda and Kajava (2009) proposed T-REKS, which uses a clustering approach based on k-means.

The systems HHrep (Soding et al., 2006) and HHRepID (Biegert and Soding, 2008) are instead based on building and matching Hidden Markov Models for the repeating substrings to be sought (not necessarily tandem).

Some approaches based on neural networks aim at detecting particular repetitive structures. For example, Palidwor et al. (2009) developed a classification technique for detecting alpharods repeats, a specific important repetitive structure [see also Rubinson and Eichman (2012)].

For the class of protein solenoid repeats, REPETITA, by Marsella et al. (2009), uses several AA biochemical properties (including polarity, secondary structure, molecular volume, electric charge, and codon diversity) and a discrete fourier transform approach to detect self-similarities.

Pellegrini et al. (2012) propose the notion of *fuzzy TR* (FTR) for proteins, which is based on using a normalized BLOSUM-weighted edit distance between AA sub-strings and in assuming that in a FTR, even if the constitutive unit elements may be pairwise at high divergence, there exists an "origin" string, not necessarily still part of the protein in exam, that is at a relatively small divergence from any of its unit elements. Here, the notion of high/low divergence is relative to the divergence between random AA strings under the chosen weighted edit distance. An exhaustive search of FTRs in long proteins is computationally demanding, since the bare definition leads to an NP-hard problem. Thus, an efficient heuristic is used in *PTRStalker* to guess the candidate "origin" strings.

Gruber et al. (2005) propose REPPER a meta searching approach that combines the output of different algorithms. A web-based meta-search server that allows to run and compare easily several tools on the same input is also described in Schaper et al. (2015).

Shapper et al. (Schaper et al., 2012; Anisimova et al., 2015) propose a statistical method based on phylogenetic fingerprints and ML-estimation that, in conjunction with one or more standard predictors, is able to filter out predicted TR that are more likely to be false-positive.

As screening large portions of protein sequence DB looking for TR patterns is time consuming, Richard and Kajava (2014)

TABLE 1 | Synthetic table of resources for PTR studies: sequence-based algorithms.

Name	Type	Year	Reference	Notes
INTREP	Alg	1999	Pellegrini et al. (1999)	http://nihserver.mbi.ucla.edu/Repeats/
prospero	Alg	1999	Mott (1999)	http://www.well.ox.ac.uk/rmott/ARIADNE/prospero.shtml
REP	Alg	2000	Andrade et al. (2000)	http://www.bork.embl.de/~andrade/papers/rep/search.html
RADAR	Alg	2000	Heger and Holm (2000)	https://github.com/AndreasHeger/radar/
REPRO	Alg	2000	George and Heringa (2000)	http://www.ibi.vu.nl/programs/reprowww/
TRUST	Alg	2004	Szklarczyk and Heringa (2004)	http://www.ibi.vu.nl/programs/trustwww/
REPPER	Alg	2005	Gruber et al. (2005)	http://toolkit.tuebingen.mpg.de/repper/
HHrep	Alg	2006	Soding et al. (2006)	http://toolkit.tuebingen.mpg.de/hhrep
TRED	Alg	2006	Sokol et al. (2007)	Available upon request
XSTREAM	Alg	2007	Newman and Cooper (2007)	http://jimcooperlab.mcdb.ucsb.edu/xstream/
HHRepID	Alg	2008	Biegert and Soding (2008)	http://toolkit.tuebingen.mpg.de/hhrepid/
ARD2	Alg	2009	Palidwor et al. (2009)	http://cbdm.mdc-berlin.de/~ard2/
T-REKS	Alg	2009	Jorda and Kajava (2009)	http://bioinfo.montp.cnrs.fr/
REPETITA	Alg	2009	Marsella et al. (2009)	http://protein.bio.unipd.it/repetita/
PTRStalker	Alg	2012	Pellegrini et al. (2012)	http://bioalgo.iit.cnr.it/
TRDistiller	Alg	2014	Richard and Kajava (2014)	Available upon request

propose a pre-screening tool (`TRDistiller`) whose purpose is to quickly filter out proteins that almost surely do not contain a TR, while retaining for further analysis the proteins carrying a TR with high probability.

As the list of possible tools to choose from becomes longer, there is an emerging need for guidance on which tool is most suitable for a given task. Unfortunately, at the best of my knowledge, no such comprehensive comparative study has been attempted yet. More limited comparative tests can be found in Pellegrini et al. (2012) where five methods (RADAR, TRUST, T-REKS, XSTREAM, and PTRStalker) are compared in their ability to detect very long PTRs ($\geq$ 4000 AA), with XSTREAM and PTRStalker emerging as the best choice for this task. A second test is aimed at detecting dimeric proteins by five tools (RADAR, TRUST, HHRep, HHRepID, and PTRStalker), with PTRStalker, TRUST, and HHRepID being able to successfully uncover such dimeric structures in some of the tested proteins. In Jorda and Kajava (2009), four methods (T-REKS, XSTREAM, `Internal Repeat Finder`, and TRED) are compared by the number of sequences they could identify as holding a PTR longer than 14 AA in the SWISSPROT database, with T-REKS giving the highest number (almost doubling the closest competitor). In Marsella et al. (2009), three methods (REPETITA, TRUST, and RADAR) are compared to assess their ability in guessing the correct periodicity in solenoid repeats, with REPETITA having an edge over the other two methods.

Protein-TR Detection Algorithms Based on Structure

Functional features are more readily linked to the structural features of a protein rather than to their primary sequence, thus available structural data should also be used to detect protein 3d symmetries and repetitive 3d motifs (Goodsell and Olson, 2000). However, only for a fraction of the known protein sequences, the corresponding 3D conformation could be determined, therefore the range of applicability of structure-based methods is limited w.r.t. the range of the sequence-based methods.

In this case, the algorithmic challenge lies in the multidimensional nature of the data, and on the fact that the space of rigid transformations (rotations, translations) as well as the inherent flexibility of proteins must be taken into account when attempting to match 3d substructures in order to detect the PTR periodicity. **Table 2** reports the list of structure-based algorithms.

In Murray et al. (2002), both the sequence and the structure signals are integrated within a continuous wavelet transform approach to detect repeating motifs. In particular, the sequence is represented by values of the Kyte–Doolittle hydrophobicity scale, while structure is characterized via the relative accessible surface area. This approach has been shown to be successful on most of the well known types of repetitive motifs.

DAVROS (Murray et al., 2004) is a PTR prediction system that builds upon a structural alignment program (SAP) that evaluates internal structural symmetries via a protein self-similarity matrix and employs a Fourier Transform approach to identify strong signals over the noisy background.

Swelfe (Abraham et al., 2008) finds internal repeats by combining three abstraction levels. Swelfe quickly identifies statistically significant internal repeats in DNA sequence, in the amino acid sequence and in the 3D structures using dynamic programing. The associated web server also shows the relationships between repeating feature at each level and facilitates visualization of the results.

SymD (Kim et al., 2010) is an algorithm that aims at detecting internal spatial symmetries of proteins. It uses the alignment method in Kim et al. (2009) on pairs of structure formed by the target protein and its shifted versions built by all circular permutations of its residues. Although not all PTR give rise to symmetric 3D structures, many do, therefore this approach often indicates the presence of a PTR. Other methods based on this symmetry detection approach are RQA (Chen et al., 2009), OPAAS (Shih and Hwang, 2004), and Gplus (Guerler et al., 2009).

ProSTRIP (Sabarinathan et al., 2010) uses dynamic programing to find similar structural repeats in a protein structure encoded by the protein backbone dihedral angles.

RAPHAEL (Walsh et al., 2012b) is a more recent method for the detection of solenoids in protein structures. It aims at mimicking the periodicity and distance patterns detection criteria a human curator is likely to exploit when assessing the presence of a solenoid visually. In particular, the candidate protein is subject to a random rotation and translation, and subsequently for each of the three C-alpha coordinates a projection is performed. This operation produces a profile curve, in which the distance between consecutive local maxima is a candidate periodicity value. By averaging over multiple random rotations and translations, a robust

TABLE 2 | Synthetic table of resources for PTR studies: structure-based algorithms.

Name	Type	Year	Reference	Notes
DAVROS	Alg	2004	Murray et al. (2004)	http://www.ebi.ac.uk/~murray/davros/
OPAAS	Alg	2004	Shih and Hwang (2004)	http://www.ibms.sinica.edu.tw/
Swelfe	Alg	2008	Abraham et al. (2008)	http://wwwabi.snv.jussieu.fr/public/Swelfe/
RQA	Alg	2009	Chen et al. (2009)	
Gplus	Alg	2009	Guerler et al. (2009)	http://agknapp.chemie.fu-berlin.de/gplus/
SymD	Alg	2010	Kim et al. (2010)	http://symd.nci.nih.gov/
ProSTRIP	Alg	2010	Sabarinathan et al. (2010)	http://cluster.physics.iisc.ernet.in/prostrip/
RAPHAEL	Alg	2012	Walsh et al. (2012b)	http://protein.bio.unipd.it/raphael/
Frustratometer	Alg	2013	Parra et al. (2013)	http://www.proteinphysiologylab.tk/
ConSole	Alg	2014	Hrabe and Godzik (2014)	http://console.sanfordburnham.org/
PRIGSA	Alg	2014	Chakrabarty and Parekh (2014)	http://bioinf.iiit.ac.in/PRIGSA/

period estimation is attained. Additional simple rules allow to further detect non-periodic residues interspersed in the solenoid periodic structure.

Parra et al. (2013) use the structural alignment tool TopMatch (Sippl, 2008) to search exhaustively the space of possible sub-structures that tile a large fraction of a given structure, and thus can represent a *bona fide* structural repetitive element of the input protein.

PRIGSA (Chakrabarty and Parekh, 2014) represents distance information among residues in an adjacency matrix, and it is based on the observation that similar sub-structures can be recognized as unique profiles of the principal eigenspectra of this matrix.

ConSole (Hrabe and Godzik, 2014) aims at detecting solenoid domains having as input structural information, by searching repetitive patterns in a *contact matrix*, which, for every pair of residues *i,j* in a protein, encodes a value 1 if the two residues have at least a pair of heavy atoms at Euclidean distance below a threshold t (set at $t = 4.5$ Å). *Ad hoc* rules are further applied in order to handle insertions in the solenoid repetitive patterns.

As in the case of sequence-based methods, very few comparative studies among the proposed structure-based tools have been done. In Kim et al. (2010), six methods (DAVROS, OPAAS, Swelfe, RQA, Gplus, and SymD) are compared in their ability to identify characteristic symmetries in fold families from CATH, SCOP, and ASTRAL databases, with SymD having an overall better performance. In Sabarinathan et al. (2010), two methods (ProSTRIP and Swelfe) are compared over well known families of repeat proteins, for the task of detecting periodicity and exact repeat positions. On well known PTR proteins, both methods detect approximatively the correct period, however, ProSTRIP detects more repeating units. On the harder class of multidomain proteins ProSTRIP is also better at guessing the correct periodicity. In Walsh et al. (2012b) five methods (both sequence and structure based) are compared (namely Swelfe, RAPHAEL, REPETITA, TRUST, and RADAR) in their ability to guess the PTR periodicity, with RAPHAEL giving better predictions, when we allow for a slackness of 5 AA in the predicted value. For exact predictions, RAPHAEL, REPETITA, and TRUST are about equivalent.

Databases for Protein-TR

Information about PTR can be retrieved as annotations in general purpose integrated protein databases. However, such annotations often cover only the well studied PTR, therefore in recent years a number of special purpose repositories have been assembled with the objective of making large scale PTR analysis easier. We list here in **Table 3** only DBs that are available on-line at the present time, as many older published articles refer to DBs no longer available.

RepSeq (Depledge et al., 2007) is a specialized DB for PTR in lower eukaryotic pathogens.

PRDB is a PTR database that supports queries on protein tandem repeats found in sequence data bases. Currently, it holds about 1.25M PTR extracted from the Swissprot, PDB, and NR databases in early 2010 using the T-REKS detection tool (Jorda et al., 2012). This database has been instrumental for uncovering original biological correlations in Jorda et al. (2010).

TABLE 3 | Synthetic table of resources for PTR studies: databases.

Name	Type	Year	Reference	Notes
RepSeq	DB	2007	Depledge et al. (2007)	http://www.repseq.org/
PRDB	DB	2012	Jorda et al. (2012)	http://bioinfo.montp.cnrs.fr/
ProRepeat	DB	2012	Luo et al. (2012)	http://prorepeat.bioinformatics.nl/
PTRStalkerDB	DB	2012	Pellegrini et al. (2012)	http://bioalgo.iit.cnr.it/
RepeatsDB	DB	2013	Di Domenico et al. (2013)	http://repeatsdb.bio.unipd.it/

PTRStalkerDB lists the PTR found with the PTRStalker method on the SwissProt database release 57.15 of March 2, 2010 that contains 515,203 sequence entries.

ProRepeat (Luo et al., 2012) is a curated and integrated data base and analysis platform for research on the biological features of amino acid tandem repeats. ProRepeat collects PTR of protein sequences listed in the UniProt knowledge base from different species; moreover, it includes 85 completely sequenced eukaryotic proteomes from the RefSeq collection. The latest datasets used in ProRepeat are UniProtKB Release May 2011 and RefSeq Release 40.

RepeatsDB (Di Domenico et al., 2013) is a database of annotated tandem repeat protein *structures* that uses both a state of the art detection method (RAPHAEL) and manual curation to survey the protein structures listed in PDB. The latest version 2.0.0 (beta) released in 2015 holds 10,039 PTR structures (including manually classified and predicted PTR). Automated updates every 3 months are planned.

Although progress in the area of databases for PTR has come about in the past few years, there is also much scope for improvement, in particular, as the amount of proteomic data increases rapidly, it is important to maintain the PTR databases aligned with the latest releases of the reference protein sequence and structure. Also, given the variety of algorithms and approaches to PTR prediction, DB that uses one single algorithm as source of data could suffer for the specific algorithm's biases, and more robust prediction could be obtained instead by using multiple detecting algorithms.

Classification of Protein-TR

Kajava (2012) reports an extensive survey of bioinformatic tools to support various analysis of TR in proteins, including tools for identification of TR in proteins, databases reporting PTR (either exclusively, or as an annotation in a larger protein DB), classification of repetitive 3D structures, and tools for structural prediction targeting proteins with PTR (as opposed to globular ones).

Early surveys by Marcotte et al. (1998), Andrade et al. (2001), and Kajava (2001) are very much concerned with the task of identifying specific classes of proteins highly characterized by their PTR content with the aim of finding corresponding structural and functional regularities. Andrade et al. (2001) propose a taxonomy of six main classes (β-propellers, β-trefoils,

TPR-like, Ankyrin-like, Armadillo/HEAT-like and Leucine-Rich). Instead Kajava (2001) uses a classification based on the repeating unit length (1–2 residues = class I crystalline aggregates, 3–4 residues = class II fibrous proteins, 5–40 residues = class III solenoid-like proteins, and class IV beads-on-string proteins with repeats longer than 30 residues folded into globular domains). Later in Kajava (2012), a refinement of this classification by splitting class III into two sub-classes of *solenoid* and *non-solenoid* structures has been proposed. The database RepeatsDB (Di Domenico et al., 2013) uses the classification proposed by Kajava (2012).

Mechanisms of Protein-TR Expansion During Evolution

Björklund et al. (2006) and Moore et al. (2008) analyze the internal sequence similarity in proteins of several species and note that the domain repeats are often expanded through simultaneous duplications of several domains in one event, while the duplication of one domain at a time is a less common event. Moreover, many of the repeats appear to have been duplicated in the middle of the repeat region. This behavior is in contrast to the evolution of other proteins that mainly happens through additions of single domains at either terminus of the protein. No common mechanism for the expansion of all repeats could be detected in this study, for example, duplication patterns show no dependence on the size of the domains. Repeat expansion in some families can possibly be explained by shuffling of exons but exon shuffling does not appear to be a general formation mechanism.

Some domain families show distinct specific duplication patterns, for example, nebulin domains have mainly been expanded with groups of seven domains at a time, while duplications of other domain families involve varying numbers of domains for each event. A more detailed analysis of nebulin domains evolution is in Björklund et al. (2010).

By mapping the Protein TR back onto their coding DNA sequences, Street et al. (2006) study the conservation of intron/exon patterns across several species and show evidence that subdivide the repeat protein genes into two classes. The first class has random-length exons that are likely produced by accumulating introns though random insertion within the array of repetitive units. The second class is composed exclusively of exons corresponding to the multiple of the repeating unit, and thus is likely to be formed by local duplications of intron/exon modules.

Protein-TR Evolutionary Conservation

In Schaper et al. (2014), it is described a proteome-wide analysis of the evolution of TR in human proteins, using a database of 61 eukaryotes. The main finding is that the vast majority of human PTR are ancient, with TR unit number and order preserved intact since remote speciation events. Moreover, no human PTR shows evidence of a recent duplication or deletion event. Thus, presumably, most PTRs fold into stable and conserved structures, indispensable for their function. Similar findings for plants are shown in Schaper and Anisimova (2015). The analysis of PTR in *Drosophila melanogaster* reported in Ponting et al. (2001) led to the identification of novel PTR in the products of disease-related human genes homologous to those in *Drosophila melanogaster*.

Protein-TR in Protein Design

Different structures which arise from tandem arrays of a repeated structural motif have generated significant interest with respect to protein engineering and synthetic protein design (Forrer et al., 2003, 2004; Main et al., 2003, 2005; Javadi and Itzhaki, 2013). Several results are reported in these articles about re-engineering of PTR binding specificities, with attention to protein folding kinetics and protein stability.

Sawyer et al. (2013) present a "module-based" design approach in which modules composed of tandem repeats are aligned to identify repeat-specific features that will be important to include in future repeat protein design templates.

Parmeggiani et al. (2015) describe a general database-driven approach for reliable generation of synthetic stable modular repeat proteins. Concomitant to the distillation of general design principles for PTR engineering, research activities have been also concentrated toward specific classes of Protein-TR which have shown a more promising potential for applications (Stumpp et al., 2015). A notable example is that of *Designed Ankyrin Repeat Proteins (DARPins)* (Binz2003) that have been extensively studied [see a recent survey by Plückthun (2015) and references therein], since they provide a biochemically stable scaffold for designing protein variants able to recognize targets with affinity and specificity that are equal or possibly superior to that of antibodies. Similar promising studies focus also on *armadillo repeat proteins* (Reichen et al., 2014) and *leucine-rich-repeat proteins* (Park et al., 2015).

Order, Disorder, and Protein-TR

While our view of protein functions is often linked to the presence of a well defined 3-dimensional protein conformations, it has been recognized (Tompa, 2002) that many important protein functions are also linked to proteins (or regions within a protein) that lack a folded structure, but display a highly flexible random-coil-like conformation under physiological conditions [named intrinsically unstructured proteins (IUP) or intrinsically unstructured regions (IUR)].

The concept of order and disorder in protein segments (Dunker et al., 2001; Tompa, 2002) has been often investigated in correlation with the presence or absence of PTR at the sequence level. For example in Tompa (2002), 21 IUP are examined, and further 21 cases are cited in Dunker et al. (2001). It is noticed that IUR often correspond to regions of low compositional complexity (low sequence entropy) and sometimes to repetitive sub-sequences in fibrillar proteins. Tompa and Fersht (2009) discuss in detail the cases of PTR in PEVK regions of human Titin, in prion proteins and in the CTD domain of RNA polymerase. These findings on specific instances are, however, hard to generalize.

A general property observed by Jorda et al. (2010) is that higher level of repeat perfection correlates positively with the disordered state of protein sub-chains.

The emergence of IUP/IUR prediction tools, such as IUpred (Dosztányi et al., 2005), ESpritz (Walsh et al., 2012a), and DISOPRED (Jones and Cozzetto, 2015), to name a few, and

comprehensive databases of IUP/IUR, such as DisProt (Sickmeier et al., 2007) and MobiDB 2.0 (Potenza et al., 2015), can be quite useful for finding generalizable connections between PTR and ordered/disordered states of protein regions.

Correlation of Protein-TR with Other Protein Properties

In Turutina et al. (2006), the sequences of the *Swiss-Prot* protein families are analyzed in order to detect family-specific latent periodicity fingerprints induced by PTR, using the method in Korotkov et al. (2003), and 94 such protein families are reported as well-characterized by such fingerprints.

A complete analysis of PDB sequences using RADAR is reported in Rajathei and Selvaraj (2013), where a good correlation among PTR, structural similarity, and functionally involved residues is highlighted.

In Mularoni et al. (2007) and Mularoni et al. (2010), the function and evolution of a particular class of PTR formed by repetitions of a *single AA* are investigated (homo-TR). These two studies concentrated on human and mouse homo-TR of length four. The protein stabilizing properties of homo-TR are also reported in Katti et al. (2000). A more general statistical analysis of homo-PTR in human proteins is in Jorda and Kajava (2010).

Conclusion

The present survey on Protein-TR touches several aspects of this research fields, including detection algorithms (Sections "Protein-TR Detection Algorithms Based on Sequence" and "Protein-TR Detection Algorithms Based on Structure"), databases (Section "Databases for Protein-TR"), classification (Section "Classification of Protein-TR"), the relationship between PTR and biologically relevant concepts (Sections "Mechanisms of Protein-TR Expansion During Evolution," "Protein-TR Evolutionary Conservation," "Order, Disorder, and Protein-TR," and "Correlation of Protein-TR with Other Protein Properties"), and it highlights also recent progress in the design of synthetic PTR (Section "Protein-TR in Protein Design").

Although there has been steady progress in the last 15 years in devising new prediction tools, both sequence and structure based, very little comparative or integrative work has been done. Most of the proteome-wise studies use only one tool to define and detect PTR and draw conclusions on PTR distributions and statistics. Though this approach was completely justified in the pioneering times (late 1990s and early 2000s), it is necessary now to refine these methodologies and make full use of the wealth of algorithms and approaches devised in the last decade. A more robust assessment of the distribution and annotations of PTR over the entire proteome could be attained by applying and merging the outcomes of multiple tools. In this context, the manually curated databases of PTRs can provide the necessary validation benchmarks.

From the point of view of the design of prediction tools, one open challenge is to devise sequence-based tools that are able to come close to the performance of structure-based tools. Thus providing higher quality PTR predictions for a larger pool of sequenced proteins.

Funding

The present work is partially supported by the Flagship project InterOmics (PB. P05), funded by the Italian Ministry for Instruction University and Research (MIUR) and CNR organizations, and by the joint IIT-IFC Laboratory of Integrative Systems Medicine (LISM).

References

Abraham, A.-L., Rocha, E. P. C., and Pothier, J. (2008). Swelfe: a detector of internal repeats in sequences and structures. *Bioinformatics* 24, 1536–1537. doi:10.1093/bioinformatics/btn234

Andrade, M. A., Perez-Iratxeta, C., and Ponting, C. P. (2001). Protein repeats: structures, functions, and evolution. *J. Struct. Biol.* 134, 117–131. doi:10.1006/jsbi.2001.4392

Andrade, M. A., Ponting, C. P., Gibson, T. J., and Bork, P. (2000). Homology-based method for identification of protein repeats using statistical significance estimates. *J. Mol. Biol.* 298, 521–537. doi:10.1006/jmbi.2000.3684

Anisimova, M., Pečerska, J., and Schaper, E. (2015). Statistical approaches to detecting and analyzing tandem repeats in genomic sequences. *Front. Bioeng. Biotechnol.* 3:31. doi:10.3389/fbioe.2015.00031

Apweiler, R., Bairoch, A., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., et al. (2004). Uniprot: the universal protein knowledgebase. *Nucleic Acids Res.* 32(Suppl. 1), D115–D119. doi:10.1093/nar/gkh131

Biegert, A., and Soding, J. (2008). De novo identification of highly diverged protein repeats by probabilistic consistency. *Bioinformatics* 24, 807–814. doi:10.1093/bioinformatics/btn039

Björklund, A. K., Light, S., Sagit, R., and Elofsson, A. (2010). Nebulin: a study of protein repeat evolution. *J. Mol. Biol.* 402, 38–51. doi:10.1016/j.jmb.2010.07.011

Björklund, ÅK., Ekman, D., and Elofsson, A. (2006). Expansion of protein domain repeats. *PLoS Comput. Biol.* 2:e114. doi:10.1371/journal.pcbi.0020114

Chakrabarty, B., and Parekh, N. (2014). Prigsa: protein repeat identification by graph spectral analysis. *J. Bioinform. Comput. Biol.* 12, 1442009. doi:10.1142/S0219720014420098

Chen, H., Huang, Y., and Xiao, Y. (2009). A simple method of identifying symmetric substructures of proteins. *Comput. Biol. Chem.* 33, 100–107. doi:10.1016/j.compbiolchem.2008.07.026

Depledge, D. P., Lower, R. P., and Smith, D. F. (2007). Repseq – a database of amino acid repeats present in lower eukaryotic pathogens. *BMC Bioinformatics* 8:122. doi:10.1186/1471-2105-8-122

Di Domenico, T., Potenza, E., Walsh, I., Gonzalo Parra, R., Giollo, M., Minervini, G., et al. (2013). Repeatsdb: a database of tandem repeat protein structures. *Nucleic Acids Res.* 42, D352–D357. doi:10.1093/nar/gkt1175

Dosztányi, Z., Csizmok, V., Tompa, P., and Simon, I. (2005). Iupred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics* 21, 3433–3434. doi:10.1093/bioinformatics/bti541

Dunker, A., Lawson, J., Brown, C. J., Williams, R. M., Romero, P., Oh, J. S., et al. (2001). Intrinsically disordered protein. *J. Mol. Graph. Model.* 19, 26–59. doi:10.1016/S1093-3263(00)00138-8

Forrer, P., Binz, H. K., Stumpp, M. T., and Plückthun, A. (2004). Consensus design of repeat proteins. *Chembiochem* 5, 183–189. doi:10.1002/cbic.200300762

Forrer, P., Stumpp, M. T., Binz, H., and Plückthun, A. (2003). A novel strategy to design binding molecules harnessing the modular nature of repeat proteins. *FEBS Lett.* 539, 2–6. doi:10.1016/S0014-5793(03)00177-7

George, R., and Heringa, J. (2000). The repro server: finding protein internal sequence repeats through the web. *Trends Biochem. Sci.* 25, 515–517. doi:10.1016/S0968-0004(00)01643-1

Goodsell, D. S., and Olson, A. J. (2000). Structural symmetry and protein function. *Annu. Rev. Biophys. Biomol. Struct.* 29, 105–153. doi:10.1146/annurev.biophys.29.1.105

Gruber, M., Soding, J., and Lupas, A. N. (2005). REPPER-repeats and their periodicities in fibrous proteins. *Nucleic Acids Res.* 33(Suppl._2), W239–W243. doi:10.1093/nar/gki405

Guerler, A., Wang, C., and Knapp, E.-W. (2009). Symmetric structures in the universe of protein folds. *J. Chem. Inf. Model.* 49, 2147–2151. doi:10.1021/ci900185z

Heger, A., and Holm, L. (2000). Rapid automatic detection and alignment of repeats in protein sequences. *Proteins* 41, 224–237. doi:10.1002/1097-0134(20001101)41:2<224::AID-PROT70>3.0.CO;2-Z

Heringa, J., and Argos, P. (1993). A method to recognize distant repeats in protein sequences. *Proteins* 17, 391–411. doi:10.1002/prot.340170407

Hrabe, T., and Godzik, A. (2014). Console: using modularity of contact maps to locate solenoid domains in protein structures. *BMC Bioinformatics* 15:119. doi:10.1186/1471-2105-15-119

Javadi, Y., and Itzhaki, L. S. (2013). Tandem-repeat proteins: regularity plus modularity equals design-ability. *Curr. Opin. Struct. Biol.* 23, 622–631. doi:10.1016/j.sbi.2013.06.011

Jones, D. T., and Cozzetto, D. (2015). Disopred3: precise disordered region predictions with annotated protein-binding activity. *Bioinformatics* 31, 857–863. doi:10.1093/bioinformatics/btu744

Jorda, J., Baudrand, T., and Kajava, A. V. (2012). Prdb: protein repeat database. *Proteomics* 12, 1333–1336. doi:10.1002/pmic.201100534

Jorda, J., and Kajava, A. V. (2009). T-REKS: identification of Tandem REpeats in sequences with a K-meanS based algorithm. *Bioinformatics* 25, 2632–2638. doi:10.1093/bioinformatics/btp482

Jorda, J., and Kajava, A. V. (2010). "Protein homorepeats: sequences, structures, evolution, and functions," in *Advances in Protein Chemistry and Structural Biology*, Vol. 79, ed. A. McPherson (Waltham, MA: Academic Press), 59–88.

Jorda, J., Xue, B., Uversky, V. N., and Kajava, A. V. (2010). Protein tandem repeats: the more perfect, the less structured. *FEBS J.* 277, 2673–2682. doi:10.1111/j.1742-4658.2010.07684.x

Kajava, A. V. (2001). Review: proteins with repeated sequencestructural prediction and modeling. *J. Struct. Biol.* 134, 132–144. doi:10.1006/jsbi.2000.4328

Kajava, A. V. (2012). Tandem repeats in proteins: from sequence to structure. *J. Struct. Biol.* 179, 279–288. doi:10.1016/j.jsb.2011.08.009

Katti, M. V., Sami-Subbu, R., Ranjekar, P. K., and Gupta, V. S. (2000). Amino acid repeat patterns in protein sequences: their diversity and structural-functional implications. *Protein Sci.* 9, 1203–1209. doi:10.1110/ps.9.6.1203

Kim, C., Basner, J., and Lee, B. (2010). Detecting internally symmetric protein structures. *BMC Bioinformatics* 11:303. doi:10.1186/1471-2105-11-303

Kim, C., Tai, C.-H., and Lee, B. (2009). Iterative refinement of structure-based sequence alignments by seed extension. *BMC Bioinformatics* 10:210. doi:10.1186/1471-2105-10-210

Korotkov, E. V., Korotkova, M. A., and Kudryashov, N. A. (2003). Information decomposition method to analyze symbolical sequences. *Phys. Lett. A* 312, 198–210. doi:10.1016/S0375-9601(03)00641-8

Luo, H., Lin, K., David, A., Nijveen, H., and Leunissen, J. A. M. (2012). Prorepeat: an integrated repository for studying amino acid tandem repeats in proteins. *Nucleic Acids Res.* 40, D394–D399. doi:10.1093/nar/gkr1019

Luo, H., and Nijveen, H. (2014). Understanding and identifying amino acid repeats. *Brief. Bioinformatics* 15, 582–591. doi:10.1093/bib/bbt003

Main, E. R., Jackson, S. E., and Regan, L. (2003). The folding and design of repeat proteins: reaching a consensus. *Curr. Opin. Struct. Biol.* 13, 482–489. doi:10.1016/S0959-440X(03)00105-2

Main, E. R., Lowe, A. R., Mochrie, S. G., Jackson, S. E., and Regan, L. (2005). A recurring theme in protein engineering: the design, stability and folding of repeat proteins. *Curr. Opin. Struct. Biol.* 15, 464–471. doi:10.1016/j.sbi.2005.07.003

Marcotte, E. M., Pellegrini, M., Yeates, T. O., and Eisenberg, D. (1998). A census of protein repeats. *J. Mol. Biol.* 293, 151–160. doi:10.1006/jmbi.1999.3136

Marsella, L., Sirocco, F., Trovato, A., Seno, F., and Tosatto, S. C. (2009). Repetita: detection and discrimination of the periodicity of protein solenoid repeats by discrete fourier transform. *Bioinformatics* 25, i289–i295. doi:10.1093/bioinformatics/btp232

Moore, A. D., Bjorklund, A. K., Ekman, D., Bornberg-Bauer, E., and Elofsson, A. (2008). Arrangements in the modular evolution of proteins. *Trends Biochem. Sci.* 33, 444–451. doi:10.1016/j.tibs.2008.05.008

Mott, R. (1999). Local sequence alignments with monotonic gap penalties. *Bioinformatics* 15, 455–462. doi:10.1093/bioinformatics/15.6.455

Mularoni, L., Ledda, A., Toll-Riera, M., and Albà, M. M. (2010). Natural selection drives the accumulation of amino acid tandem repeats in human proteins. *Genome Res.* 20, 745–754. doi:10.1101/gr.101261.109

Mularoni, L., Veitia, R. A., and Albà, M. M. (2007). Highly constrained proteins contain an unexpectedly large number of amino acid tandem repeats. *Genomics* 89, 316–325. doi:10.1016/j.ygeno.2006.11.011

Murray, K. B., Gorse, D., and Thornton, J. M. (2002). Wavelet transforms for the characterization and detection of repeating motifs. *J. Mol. Biol.* 316, 341–363. doi:10.1006/jmbi.2001.5332

Murray, K. B., Taylor, W. R., and Thornton, J. M. (2004). Toward the detection and validation of repeats in protein structure. *Proteins* 57, 365–380. doi:10.1002/prot.20202

Newman, A., and Cooper, J. (2007). Xstream: a practical algorithm for identification and architecture modeling of tandem repeats in protein sequences. *BMC Bioinformatics* 8:382. doi:10.1186/1471-2105-8-382

Palidwor, G. A., Shcherbinin, S., Huska, M. R., Rasko, T., Stelzl, U., Arumughan, A., et al. (2009). Detection of alpha-rod protein repeats using a neural network and application to huntingtin. *PLoS Comput. Biol.* 5:e1000304. doi:10.1371/journal.pcbi.1000304

Park, K., Shen, B. W., Parmeggiani, F., Huang, P.-S., Stoddard, B. L., and Baker, D. (2015). Control of repeat-protein curvature by computational protein design. *Nat. Struct. Mol. Biol.* 22, 167–174. doi:10.1038/nsmb.2938

Parmeggiani, F., Huang, P.-S., Vorobiev, S., Xiao, R., Park, K., Caprari, S., et al. (2015). A general computational approach for repeat protein design. *J. Mol. Biol.* 427, 563–575. doi:10.1016/j.jmb.2014.11.005

Parra, R. G., Espada, R., Sánchez, I. E., Sippl, M. J., and Ferreiro, D. U. (2013). Detecting repetitions and periodicities in proteins by tiling the structural space. *J. Phys. Chem. B* 117, 12887–12897. doi:10.1021/jp402105j

Pellegrini, M., Marcotte, E. M., and Yeates, T. O. (1999). A fast algorithm for genome-wide analysis of proteins with repeated sequences. *Proteins* 35, 440–446. doi:10.1002/(SICI)1097-0134(19990601)35:4<440::AID-PROT7>3.0.CO;2-Y

Pellegrini, M., Renda, M. E., and Vecchio, A. (2012). Ab initio detection of fuzzy amino acid tandem repeats in protein sequences. *BMC Bioinformatics* 13(Suppl. 3):S8. doi:10.1186/1471-2105-13-S3-S8

Plückthun, A. (2015). Designed ankyrin repeat proteins (darpins): binding proteins for research, diagnostics, and therapy. *Annu. Rev. Pharmacol. Toxicol.* 55, 489–511. doi:10.1146/annurev-pharmtox-010611-134654

Ponting, C. P., Mott, R., Bork, P., and Copley, R. R. (2001). Novel protein domains and repeats in drosophila melanogaster: insights into structure, function, and evolution. *Genome Res.* 11, 1996–2008. doi:10.1101/gr.198701

Potenza, E., Domenico, T. D., Walsh, I., and Tosatto, S. C. E. (2015). Mobidb 2.0: an improved database of intrinsically disordered and mobile proteins. *Nucleic Acids Res.* 43, 315–320. doi:10.1093/nar/gku982

Rajathei, D. M., and Selvaraj, S. (2013). Analysis of sequence repeats of proteins in the {PDB}. *Comput. Biol. Chem.* 47, 156–166. doi:10.1016/j.compbiolchem.2013.09.001

Reichen, C., Madhurantakam, C., Plückthun, A., and Mittl, P. R. (2014). Crystal structures of designed armadillo repeat proteins: implications of construct design and crystallization conditions on overall structure. *Protein Sci.* 23, 1572–1583. doi:10.1002/pro.2535

Richard, F. D., and Kajava, A. V. (2014). Trdistiller: a rapid filter for enrichment of sequence datasets with proteins containing tandem repeats. *J. Struct. Biol.* 186, 386–391. doi:10.1016/j.jsb.2014.03.013

Rubinson, E. H., and Eichman, B. F. (2012). Nucleic acid recognition by tandem helical repeats. *Curr. Opin. Struct. Biol.* 22, 101–109. doi:10.1016/j.sbi.2011.11.005

Sabarinathan, R., Basu, R., and Sekar, K. (2010). Prostrip: a method to find similar structural repeats in three-dimensional protein structures. *Comput. Biol. Chem.* 34, 126–130. doi:10.1016/j.compbiolchem.2010.03.006

Sawyer, N., Chen, J., and Regan, L. (2013). All repeats are not equal: a module-based approach to guide repeat protein design. *J. Mol. Biol.* 425, 1826–1838. doi:10.1016/j.jmb.2013.02.013

Schaper, E., and Anisimova, M. (2015). The evolution and function of protein tandem repeats in plants. *New Phytol.* 206, 397–410. doi:10.1111/nph.13184

Schaper, E., Gascuel, O., and Anisimova, M. (2014). Deep conservation of human protein tandem repeats within the eukaryotes. *Mol. Biol. Evol.* 31, 1132–1148. doi:10.1093/molbev/msu062

Schaper, E., Kajava, A. V., Hauser, A., and Anisimova, M. (2012). Repeat or not repeat? Statistical validation of tandem repeat prediction in genomic sequences. *Nucleic Acids Res.* 40, 10005–10017. doi:10.1093/nar/gks726

Schaper, E., Korsunsky, A., Messina, A., Murri, R., Pečerska, J., Stockinger, H., et al. (2015). Tral: tandem repeat annotation library. *Bioinformatics* 31, 3051–3053. doi:10.1093/bioinformatics/btv306

Shih, E. S., and Hwang, M.-J. (2004). Alternative alignments from comparison of protein structures. *Proteins* 56, 519–527. doi:10.1002/prot.20124

Sickmeier, M., Hamilton, J. A., LeGall, T., Vacic, V., Cortese, M. S., Tantos, A., et al. (2007). Disprot: the database of disordered proteins. *Nucleic Acids Res.* 35(Suppl. 1), D786–D793. doi:10.1093/nar/gkl893

Sippl, M. J. (2008). On distance and similarity in fold space. *Bioinformatics* 24, 872–873. doi:10.1093/bioinformatics/btn040

Soding, J., Remmert, M., and Biegert, A. (2006). HHrep: de novo protein repeat detection and the origin of TIM barrels. *Nucleic Acids Res.* 34(Suppl._2), W137–W142. doi:10.1093/nar/gkl130

Sokol, D., Benson, G., and Tojeira, J. (2007). Tandem repeats over the edit distance. *Bioinformatics* 23, e30–e35. doi:10.1093/bioinformatics/btl309

Street, T. O., Rose, G. D., and Barrick, D. (2006). The role of introns in repeat protein gene formation. *J. Mol. Biol.* 360, 258–266. doi:10.1093/bioinformatics/btl309

Stumpp, M. T., Forrer, P., Binz, H. K., and Pluckthun, A. (2015). *Repeat Protein from Collection of Repeat Proteins Comprising Repeat Modules*. US Patent 9,006,389.

Szklarczyk, R., and Heringa, J. (2004). Tracking repeats using significance and transitivity. *Bioinformatics* 20(Suppl. 1), i311–i317. doi:10.1093/bioinformatics/bth911

Tompa, P. (2002). Intrinsically unstructured proteins. *Trends Biochem. Sci.* 27, 527–533. doi:10.1016/S0968-0004(02)02169-2

Tompa, P., and Fersht, A. (2009). *Structure and Function of Intrinsically Disordered Proteins*. Baca Raton, FL: Chapman and Hall/CRC.

Turutina, V. P., Laskin, A. A., Kudryashov, N. A., Skryabin, K. G., and Korotkov, E. V. (2006). Identification of amino acid latent periodicity within 94 protein families. *J. Comput. Biol.* 13, 946–964. doi:10.1089/cmb.2006.13.946

Walsh, I., Martin, A. J. M., Di Domenico, T., and Tosatto, S. C. E. (2012a). Espritz: accurate and fast prediction of protein disorder. *Bioinformatics* 28, 503–509. doi:10.1093/bioinformatics/btr682

Walsh, I., Sirocco, F. G., Minervini, G., Di Domenico, T., Ferrari, C., and Tosatto, S. C. E. (2012b). Raphael: recognition, periodicity and insertion assignment of solenoid protein structures. *Bioinformatics* 28, 3257–3264. doi:10.1093/bioinformatics/bts550

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.