Syst Synth Biol (2015) 9:135–140
DOI 10.1007/s11693-015-9172-z

RESEARCH ARTICLE

Predicting stable functional peptides from the intergenic space of *E. coli*

Vipin Thomas[1] · Navya Raj[1] · Deepthi Varughese[1] · Naveen Kumar[2] · Seema Sehrawat[2] · Abhinav Grover[3] · Shailja Singh[2] · Pawan K. Dhar[1,3] · Achuthsankar S. Nair[1]

Received: 14 May 2015 / Accepted: 18 May 2015 / Published online: 29 May 2015

Abstract Expression of synthetic proteins from intergenic regions of *E. coli* and their functional association was recently demonstrated (Dhar et al. in J Biol Eng 3:2, 2009. doi:10.1186/1754-1611-3-2). This gave birth to the question: if one can make 'user-defined' genes from non-coding genome—how big is the artificially translatable genome? (Dinger et al. in PLoS Comput Biol 4, 2008; Frith et al. in RNA Biol 3(1):40–48, 2006a; Frith et al. in PLoS Genet 2(4):e52, 2006b). To answer this question, we performed a bioinformatics study of all reported *E. coli* intergenic sequences, in search of novel peptides and proteins, unexpressed by nature. Overall, 2500 *E. coli* intergenic sequences were computationally translated into 'protein sequence equivalents' and matched against all known proteins. Sequences that did not show any resemblance were used for building a comprehensive profile in terms of their structure, function, localization, interactions, stability so on. A total of 362 protein sequences showed evidence of stable tertiary conformations encoded by the intergenic sequences of *E. coli* genome. Experimental studies are underway to confirm some of the key predictions. This study points to a vast untapped repository of functional molecules lying undiscovered in the non-expressed genome of various organisms.

Keywords Intergenic sequences · Structure prediction · Functional annotation · Antimicrobial peptides

Introduction

Several studies have reported the presence of previously uncharacterized functional peptides from the 'non-coding' regions of genomes—from bacteria to human (Dinger et al. 2008; Frith et al. 2006a, b). Many of these short peptides seem to be involved in critical biological processes like signal transduction, cellular communication, metabolism, innate immunity and anti-microbial activities.

Recent reports indicate large numbers of non-coding transcripts <100 codon length ORFs encoding small peptides (Dinger et al. 2008; Frith et al. 2006a, b; Kondo et al. 2010; Kageyama et al. 2011). It is quite possible that cells may be translating some of these ncRNAs into functional peptides of <50 amino acids (Frith et al. 2006a, b). As an example, peptides encoded by a short ORF gene *pri* previously thought to encode ncRNA, have been found to control epidermal differentiation in Drosophila melanogaster (Kondo et al. 2010). Similarly, the plant gene ENOD40, previously thought to encode a non-coding RNA has been found to encode two functional peptides (Kageyama et al. 2011).

A number of short peptides (<100 amino acids) have been reported in mammalian genomes (Frith et al. 2006a, b). Computational prediction of functional domains in these proteins points towards cytochrome c-oxidase and chemokine activity. The finding that a large proportion of already known short peptides are extra-cellular indicates that these may be involved in cell–cell interactions. Interestingly, the short protein coding ORFs seem to be more evolutionarily conserved than DNA sequences.

✉ Achuthsankar S. Nair
sankar.achuth@gmail.com

1 Department of Computational Biology and Bioinformatics, University of Kerala, Thiruvananthapuram 695581, Kerala, India

2 Department of Life Sciences, School of Natural Sciences, Shiv Nadar University, Greater Noida 201316, U.P., India

3 School of Biotechnology, Jawaharlal Nehru University, New Delhi 110067, India

The recently discovered ability to artificially express non-coding sequences (Dhar et al. 2009) has given rise to sheer scale of making user-defined genes. Authors synthesized novel proteins from intergenic regions of *E. coli* that did not show any natural history of transcription. On functional evaluation, one of these proteins showed evidence of cell growth inhibition. The present study is an extension of this report with an aim to find novel protein/peptide structures from the non-coding genome which can play functional roles.

Methodology

Identifying the study sample

Intergenic sequences of *E. coli* were downloaded from the EcoGene 2.0 database. The database stores 2500 intergenic sequences along with their coordinates—proximal and distal genes, their orientation, length and so on. All the sequences were computationally translated and matched against NCBI non-redundant protein database, using BlastX tool. Of 2500 sequences, 1579 sequences were discarded from this study as their sequence similarity scores ranged over 40 %. The selected dataset of 921 sequences that showed <40 % local similarity with known proteins, were translated into six reading frames using Expasy Translate tool. We used 30 amino acids residues and above cut off to narrow down dataset of 5526 (921*6) sequences to 892 candidate sequences for making proteins.

Structure prediction

Homology based approaches could not be applied as the study was limited to sequences that lacks complete similarity with known proteins. Hence we used I-TASSER, an automated protein structure and function prediction server which combines threading and ab initio approaches for structure prediction (Zhang 2008). I-TASSER is one of the best tool available for protein structure prediction which ranked 1st in CASP7, CASP8 and CASP9 experiments. The major scoring function C-score or confidence score is an estimate of quality of predicted conformations. The C-score is usually found to span between −5 to 2 and a score closer to 2 signifies higher confidence level for the predicted protein structure. All the 892 sequences were subjected to structure prediction using I-TASSER and out of these 362 sequences found to form folded tertiary conformations with optimal C-scores ranging from −2 to 5 and formed the core data set for our study.

Stability analysis

In order to analyze the structural stability of the predicted proteins, non-bonded interactions like hydrogen bonds, hydrophobic interactions, salt bridges and disulphide bridges were computed using What If Server (Vriend 1990) and Protein Interaction Calculator (Tina et al. 2007). SCide program (Dosztanyi et al. 2003) was used for calculating stabilization centres while instability index (for estimating protein's stability in vitro) was calculated using Expasy ProtParam Tool (Gasteiger et al. 2005). The total energy of the structures was estimated in Steepest Descent method implemented in Deep View (Guex and Peitsch 1997). Cation pi interactions were found out using CaPTURE program (Gallivan and Dougherty 1999).

Sequence based functional annotations

The primary functional annotations were carried out using Protfun 2.2 server (Jensen et al. 2002). The server allows predicting the Functional and Gene Ontology categories of input amino acid sequences. The method implemented in Protfun relies on physico-chemical parameters of the amino acid sequences, protein sorting signals, and post translational modification sites, unlike other sequence homology based approaches. It was best suited in our case as the peptides are selected based on lack of complete sequence similarity with known proteins. The functional categories in Protfun predictions are Signal transducer, receptor, hormone, structural protein, transporter, ion channel, voltage-gated ion channel, transcription, transcription regulation, stress response, immune response and growth factor.

The subcellular localization of protein molecules is an important parameter that is used in annotation of gene products and in predicting potential functions (Cherian and Nair 2010). All the 362 peptide sequences were subjected to localization prediction using CELLO v.2.5 webserver. CELLO (Yu et al. 2006) is a subcellular localization prediction system based on the physicochemical parameters of amino acids that predicts whether a protein is cytoplasmic, inner membrane, periplasmic, outer membrane or extracellular with a prediction accuracy of 88.7 % for gram negative bacteria.

Results

Validating the dataset

Of 2500 intergenic sequences collected from EcoGene Database, 921 intergenic sequences (ranging from 97 to

960 nucleotides) were predicted to be non-homologous to all known proteins based on sequence similarity search against non-redundant protein database (<40 % global similarity). After translating (a) all these sequences into six reading frames and (b) applying a 30 amino acid cut off, 892 sequences without stop codons were obtained (Table 1).

Structure prediction

Of the 892 structure predicted sequences, 362 protein sequences showed evidence of a stable tertiary conformation (confidence score −2 to −5), which is a good indicator. Presence of secondary structure elements—alpha helices and beta strands (with involvement of at least five amino acids) were identified in 155 structures. Considering that most of the peptides remained within a length range of 30–50 amino acids, presence of secondary structure elements indicate conformational stability and functional potential. 132 proteins showed alpha helices, 15 showed beta strands while 8 proteins have both the secondary structure elements. 86 proteins have traces of alpha helices in their tertiary conformations, but major part was formed of loops (Fig. 1; Table 2).

52 peptides from this dataset showed linear helical structure with a considerable distribution of positively charged and hydrophobic amino acids indicating possible functional implications.

Stability analysis

Peptide structures were analyzed for the presence hydrogen bonds, hydrophobic interactions, disulphide bridges, salt bridges, stabilization centres, cation pi interactions, instability index and total energy to get an idea of their possible stability. Negative energy values, <40 instability index, higher numbers of hydrogen bonds and hydrophobic interactions, presence of stabilization centres, cation pi interactions, salt bridges and cysteine bridges are good indicators of conformational stability (Ramanathan et al. 2011).

Calculation of total energy which indicates deviation from ideal molecular parameters identified 198 PSP proteins out of 362 shows negative energy values. Instability Index calculation with ProtParam tool has identified 137 peptides with instability index lesser than 40 indicating thermodynamic stability. Variation in the number of hydrogen bonds was marked with PSP101 showing highest (84) and PSP388 with just 7 hydrogen bonds. 252 peptides shown to have >30 hydrogen bonds. Considering the size of our peptides, presence of >30 h-honds in could substantially contribute to structure integrity (Ramanathan et al. 2011). Of the 362 structures, 269 showed >10 hydrophobic interactions. Long range interactions like disulphide bridges and salt bridges were found in 51 peptides and 184 respectively. 152 peptides showed stabilization centres based on SCide predictions and 110 showed cation pi interactions (Tables 3, 4, 5).

Table 1 Length distribution of protein sequences

No. of amino acid residues	30–40	40–50	50–60	>60
No. of sequences	624	198	60	10

Majority of the sequences (69.9 % of sequences) in the dataset were 30–40 amino acids long

Sequence based function predictions

Gene Ontology predictions using Protfun 2.2 server showed a wide range of gene ontology categories for our peptide dataset. The various gene ontology categories attributed to our peptide dataset are immune response, structural protein, transcriptional regulation and signal transducer. Other major predicted categories include transcription, stress response and receptor (Table 6).

All the 362 peptide sequences were subjected to subcellular localization prediction using CELLO v.2.5 webserver (Yu et al. 2006). 282 peptides were predicted to be localized in the cytoplasm while 63 peptides are periplasmic. 12 of the dataset are characterized as extracellular while two predicted to be inner membrane peptides (Table 7).

Discussion

Functional genomics aims to characterize all functional units transcribed from the genome. Several recent studies, especially tiling array experiments suggest that the annotations so far missed a large proportion of transcripts; both protein coding and non-coding functional RNAs. It implies that many genomic regions, considered intergenic or ‘junk’ may actually encode unannotated functional products involved in various biological processes. It has been also demonstrated that many transcripts which were considered non-coding are in fact being translated to small proteins and peptides. Moreover Dhar et al. has recently described an appropriate method for the functional validation on intergenic regions.

Motivated by these recent studies, the present work aimed at identifying potential peptide structures embedded in the intergenic regions of *E. coli* that might be already expressing but missed out in annotation procedures. Moreover there may structures that were functional in ancestral organisms, or even entirely novel structures with

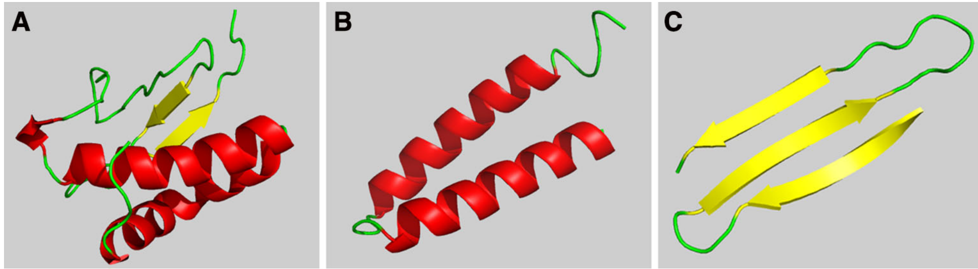

Fig. 1 Predicted structures of PSP proteins 101 (**a**), 164 (**b**) and 535 (**c**)

Table 2 Distribution of secondary structure elements in 362 structures

Secondary structure elements	Helices	B-sheets	Helix and sheets	Trace helices	Loops
No. of secondary structure elements	132	15	8	86	121

155 sequences formed true secondary structure elements with the involvement of five or more amino acids

Table 3 A sample of peptides showing stability parameters and values

Id	Length	H-bonds	Hydrophobic interactions	Disulphide bridges	Salt bridges	Stabilization centres	Cation pi	Instability index	Total energy
PSP186	48	51	36	1	1	10	1	28.85	−1365.216
PSP245	44	41	22	0	1	7	3	24.36	−941.149
PSP325	42	42	21	1	1	7	3	37.02	−1359.283
PSP432	38	53	26	1	2	2	1	33.25	−1569.342
PSP565	35	61	37	0	1	5	2	29.51	−1083.816

Table 4 Distribution of Hydrogen bonds and hydrophobic interactions

Numbers	<10	10–20	20–30	30–40	40–50	50–60	60–70	>70
H-bonds	2	15	82	131	87	33	9	1
Hydrophobic interactions	71	174	89	23	2	0	2	1

Table 5 Distribution of salt bridges and stabilization centres

Numbers	0	1	2	3	4	5	6	7	8	9	10	>10
Salt bridges	131	119	57	25	11	14	4	1	–	–	–	–
Stabilization centres	170	0	50	9	23	11	16	11	6	13	11	21

potential to perform new biological roles. In our work we have concentrated only on novel structures, hence we excluded all interegnic sequences that have a considerable sequence similarity, which yielded 892 sequences.

The specific function of a protein/peptide is closely related to its unique three dimensional globular structures. Hence one of the prerequisites for a sequence to form a functional protein is its ability to form a considerably stable tertiary conformation. Moreover, information on the structure will also provide valuable clues about the possible functions of the protein sequence under consideration. So we wanted to investigate whether the 'non-coding' sequences of our dataset can form folded tertiary structures. *Ab initio* structure prediction using I-TASSER webserver showed that 362 sequences of our dataset can form optimal tertiary conformations with the C-score ranging from −2 to 5.

Further analysis of the predicted structures shows that out of the 362 peptide structures 155 has true secondary structure elements formed of alpha helices and beta strands. This itself indicates that at least a few or these molecules can play biological roles if expressed, either naturally or artificially. Out of these 155, 132 peptides are formed mainly of alpha helices, with 32 having long linear helices. Presence of alpha helices could be an important indicator

Table 6 Predicted gene ontology categories of PSP proteins

Gene ontology	Nos.
Signal transducer	47
Receptor	33
Hormone	7
Structural protein	63
Transporter	10
Ion channel	4
Transcription	24
Transcription regulation	57
Stress response	24
Immune response	83
Growth factor	10
	362

Table 7 Sequence based subcellular predictions of PSP proteins

Subcellular location	Nos.
Inner membrane	2
Periplasmic	63
Cytoplasmic	282
Extracellular	12
Unknown	3

of potential biological activity. Many proteins exert their biological function with the help of an alpha helix segment that interact with other proteins or the DNA itself. Moreover most of the cationic peptides with anti-microbial activity are helical peptides (Harrison et al. 2010; Powers and Hancock 2003). One of the major functions attributed to short peptides with helicity is immune response and hence the anti-microbial potential of CSSB peptides has to be further investigated.

Since structure stability is an important indicator of functional potential of bioactive peptides, we adopted a computational approach to investigate certain parameters that are critical in determining structure stability. Various kinds of atomic contacts within the protein and with the medium are instrumental in protein structure formation. While short interactions like H-bonds between neighbouring amino acids involved in folding, long range interactions like disulphide bridges contribute to tertiary structure stability. The major non-covalent interactions that contribute to structure stability include H-bonds, hydrophobic interactions, disulphide bridges and salt bridges. Calculation of H-bonds, hydrophobic interactions, stabilization centres, instability index etc. indicates that a major proportion of our peptides dataset can be thermodynamically stable, further confirming their potential to play biological roles.

The primary functional annotations were carried our using Protfun 2.2 server. The server allows predicting the Functional and Gene Ontology categories of input amino acid sequences. Gene Ontology (GO) refers to a hierarchical set of terms that describe protein functions at different levels. The method implemented in Protfun relies on physico-chemical parameters of the amino acid sequences, protein sorting signals, and post translational modification sites, unlike other sequence homology based approaches. It was best suited in our case as the peptides are selected based on lack of complete sequence similarity with known proteins. Gene Ontology predictions using Protfun 2.2 server showed a wide range of gene ontology categories for our peptide dataset. The highest number of peptides (83) was attributed to in Immune response while 57 predicted as involved in Transcription regulation. As per the findings, a substantial number of peptides (47) can potentially act as signal transducers while 24 could be involved in Stress response, if expressed. This is an important observation as these four are the major functional roles attributed to short peptides in recent studies.

Conclusion

In the present study we have demonstrated a method to computationally characterize novel peptide structures from the intergenic space. With several recent studies suggesting the presence of short peptides embedded in the 'dark matter' genome of many organisms, with roles in immune response, signal transduction etc. Our study aimed at exploring the functional potential of these 'non-natural' peptides, which offers a huge untapped repository of proteins with novel functions. The prediction of 362 peptide structures with considerable conformational stability indicates that at least a few of them may be already expressed as functional peptides in *E. coli*. Further sequence based and structure based predictive methods are to be applied to narrow down the functional roles attributed to the peptides. Moreover the overwhelming presence of helical peptides augments their potential as therapeutic agents which are to be further investigated.

Acknowledgments We sincerely thank the State Inter-University Centre of Excellence in Bioinformatics (SIUCEB), University of Kerala for the funding provided during this work.

Conflict of interest The authors declare that they have no conflict of interests.

References

Cherian BS, Nair AS (2010) Protein location prediction using atomic composition and global features of the amino acid sequence. Biochem Biophys Res Commun 391:1670–1674

Dhar PK, Thwin CS, Tun K, Tsumoto Y, Maurer-Stroh S, Eisenhaber F, Surana U (2009) Synthesizing non-natural parts from natural genomic template. J Biol Eng 3:2. doi:10.1186/1754-1611-3-2

Dinger ME, Pang KC, Mercer TR, Mattick JS (2008) Differentiating protein-coding and noncoding RNA: challenges and ambiguities. PLoS Comput Biol 4(11):e1000176

Dosztanyi Z, Magyar C, Tusnady G, Simon I (2003) SCide: identification of stabilization centers in proteins. Bioinformatics 19:899–900

Frith MC, Bailey TL, Kasukawa T, Mignone F, Kummerfeld SK, Madera M, Sunkara S et al (2006a) Discrimination of non-protein-coding transcripts from protein-coding mRNA. RNA Biol 3(1):40–48

Frith MC, Forrest AR, Nourbakhsh E, Pang KC, Kai C, Kawai J, Carninci P et al (2006b) The abundance of short proteins in the mammalian proteome. PLoS Genet 2(4):e52

Gallivan JP, Dougherty DA (1999) Cation-pi interactions in structural biology. Proc Natl Acad Sci USA 96:9459

Gasteiger E, Hoogland C, Gattiker A, Duvaud S, Wilkins MR, Appel RD, Bairoch A (2005) Protein identification and analysis tools on the ExPASy server. The Proteomics Protocols Handbook. Humana Press, New York, pp 571–607

Guex N, Peitsch MC (1997) SWISS-MODEL and the Swiss-Pdb viewer: an environment for comparative protein modeling. Electrophoresis 18:2714–2723

Harrison RS, Shepherd NE, Hoang HN, Ruiz-Gómez G, Hill TA, Driver RW, Desai VS et al (2010) Downsizing human, bacterial, and viral proteins to short water-stable alpha helices that maintain biological potency. Proc Natl Acad Sci USA 107(26):11686–11691

Jensen LJ, Gupta R, Blom N, Devos D, Tamames J, Kesmir C, Nielsen H, Stærfeldt HH, Rapacki K, Workman C, Andersen CAF, Knudsen S, Krogh A, Valencia A, Brunak S (2002) Ab initio prediction of human orphan protein function from post-translational modifications and localization features. J Mol Biol 319:1257–1265

Kageyama Y, Kondo T, Hashimoto Y (2011) Coding vs non-coding: translatability of short ORFs found in putative non-coding transcripts. Biochimie 93:1981–1986

Kondo T, Plaza S, Zanet J, Benrabah E, Valenti P, Hashimoto Y, Kobayashi S, Payre F, Kageyama Y (2010) Small peptides switch the transcriptional activity of Shavenbaby during *Drosophila embryogenesis*. Science 5989:336–339

Powers J-PS, Hancock REW (2003) The relationship between peptide structure and antibacterial activity. Peptides 24:1681–1691

Ramanathan K, Shanthi V, Rajasekaran R, Sudandiradoss C, Doss CGP, Sethumadhavan R (2011) Predicting therapeutic template by evaluating the structural stability of anti-cancer peptides: a computational approach. Int J Pept Res Ther 17(1):31–38

Tina KG, Bhadra R, Srinivasan N (2007) PIC: protein interactions calculator. Nucl Acids Res 35:W473–W476

Vriend G (1990) WHAT IF: a molecular modeling and drug design program. J Mol Graph 8:52–56

Yu CS, Chen YC, Lu CH, Hwang JK (2006) Prediction of protein subcellular localization. Prot Struct Funct Bioinform 64:643–651

Zhang Y (2008) I-TASSER server for protein 3D structure prediction. BMC Bioinform 9:40