An algorithm to find all palindromic sequences in proteins

N Prasanth, M Kirti Vaishnavi and K Sekar*

Supercomputer Education and Research Centre,
Indian Institute of Science,
Bangalore 560 012

**Corresponding author (Fax, +91-080-23600683/23600551; Email, sekar@physics.iisc.ernet.in; sekar@serc.iisc.ernet.in)*

A palindrome is a set of characters that reads the same forwards and backwards. Since the discovery of palindromic peptide sequences two decades ago, little effort has been made to understand its structural, functional and evolutionary significance. Therefore, in view of this, an algorithm has been developed to identify all perfect palindromes (excluding the palindromic subset and tandem repeats) in a single protein sequence. The proposed algorithm does not impose any restriction on the number of residues to be given in the input sequence. This avant-garde algorithm will aid in the identification of palindromic peptide sequences of varying lengths in a single protein sequence.

[Prasanth N, Vaishnavi MK and Sekar K 2013 An algorithm to find all palindromic sequences in proteins. *J. Biosci.* **38** 173–177] **DOI** 10.1007/s12038-013-9300-4

1. Introduction

A palindrome is a symmetric set of characters that reads the same when read from left-to-right and right-to-left. The history of palindromes dates back to the Latin word square 'Sator Arepo Tenet Opera Rotas', which is an example of a perfect palindrome. This phenomenon can also be observed in DNA, RNA and protein sequences. A palindromic DNA sequence is a sequence with an axis of a twofold symmetry that results in the same RNA molecule when transcribed from both directions (Engberg and Klenow 1977). These sequences are often associated with (i) restriction enzyme recognition sites (for example, the palindromic sequence 'GAATTC' is recognized by the restriction enzyme EcoRI) (Roulland-Dussoix and Boyer 1969), (ii) ribosomal RNA coding genes (Engberg *et al*. 1976; Karrer and Gall 1976; Vogt and Braun 1976) and (iii) in the formation of hairpin loops in the newly transcribed RNA. Palindromic sequences are observed in various classes of proteins like histones (Cheng *et al.* 1989), prion proteins (Sulkowski 1992; Kazim 1993), DNA-binding proteins (Suzuki 1992; Ohno 1993; Giel-Pietraszuk *et al*. 2003), Rhodopsin family (Ohno 1990), metal-binding proteins (Pan *et al*. 1999), sugar-metabolizing proteins (Ohno 1992) and receptors (Jaseja *et al*. 2005). In proteins, short palindromic sequences occur more frequently than long palindromic sequences (Hoffmann and Rychlewski 1999), which are generally composed of amino acids like tryptophan, cysteine, histidine and methionine (Ohno 1992). A comprehensive study (Ohno 1992) shows that histones are rich in palindromes (>50%), whereas sugar-metabolizing proteins exhibit low composition of palindromes (<18%). Sulkowski (1992) in one short communication revealed that prion proteins contain 'an unusual peptide segment that can be designated as an aromatic palindrome' (-R-Y-Y-x-x-N-x-Y-R-Y-x-N-x-x-Y-Y-R-), which is composed of tyrosine (Y), arginine (R) and asparagine (N) residues. This palindrome can constitute the interface of a prion protein dimer and could also be involved in the binding of glial fibrillary acidic protein.

The structural conformation adopted by a palindromic sequence remains a debate because of the contradicting results available in the literature (Lacroix *et al*. 1998; Mittl *et al*. 2000; Shukla *et al*. 2003; Rai 2007; Pal-Bhowmick *et al*. 2007). A study conducted by Sheari Sheari *et al*. (2008) showed that palindromic sequences in proteins occur more frequently in low-complexity regions and have a

Keywords. Algorithm; histones; palindrome; proteins

Published online: 20 January 2013

high tendency to form α-helices. In spite of the extensive work done on palindromic peptide sequences, information regarding their structure, function and evolution is still unknown. Further, to the best knowledge of the authors, there exists no efficient method for the identification of palindromes in protein sequences. Therefore, an efficient algorithm has been developed for the identification of palindromes in protein sequences.

2. Proposed algorithm

The proposed algorithm to find all perfect palindromic sequences in a single protein sequence has been discussed below. The algorithm can be divided into two parts: (i) finding both odd- and even-length palindromes and (ii) removal of all sub-palindromes.

2.1 *Finding both odd- and even-length palindromes*

First, the input sequence is stored in a string variable 'seq' and then linear iteration is performed to obtain the palindromes. This part of the algorithm is where all the possible palindromes (both odd and even lengths) are extracted from the sequence. The element at i-th position is compared with its succeeding (i+1)th element. If they are the same, then the (i-1)th element is compared with the (i+2)th element. This iteration continues as long as the condition is satisfied, and upon encountering a mismatch, the algorithm exits the loop and the comparison is stopped. Now, the current extreme positions are paired and stored as start and end positions of the sequence. These paired positions are appended to a vector which stores the paired positions of all palindromes. Subsequently, the element at the i-th position is compared with the element at the (i+2)th position, for similarity. If they are the same, then the above iteration process is repeated and the paired positions are appended to the vector. Thus, the vector now contains all the paired positions of the palindromes.

```
while(i − k − 1 >= 0 && i + k < n &&
seq[i − k − 1]== seq[i + k]) + +k;
if(k) palins.push_back(make_pair(i − k, −(i + k − 1)));
```

For the sequence ABBSBBAB (figure 1), there would be four pairs of positions ([1, −2], [4, − 5], [0, −6], [5, −7]) present in the vector 'res'. The positions of the pairs after sorting the array would be [0, −6], [1, −2], [4, −5], [5, −7]. Adding a negative sign for the end position makes sorting easy, because the sort function arranges the positions in an ascending order. For example, for a pair like (0, 4) and (0, 8), the sort function would result in (0, 4) and then (0, 8), this would be tough for elimination. However, when the positions are (0, −4) and (0, −8), the final result would be (0, −8) and (0, −4), which helps when eliminating the sub-palindromes.

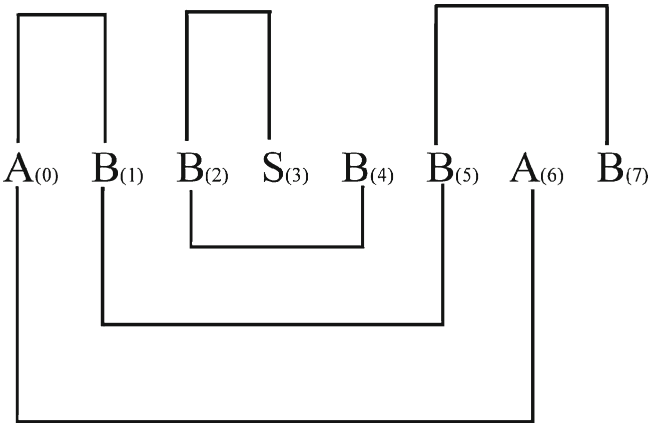
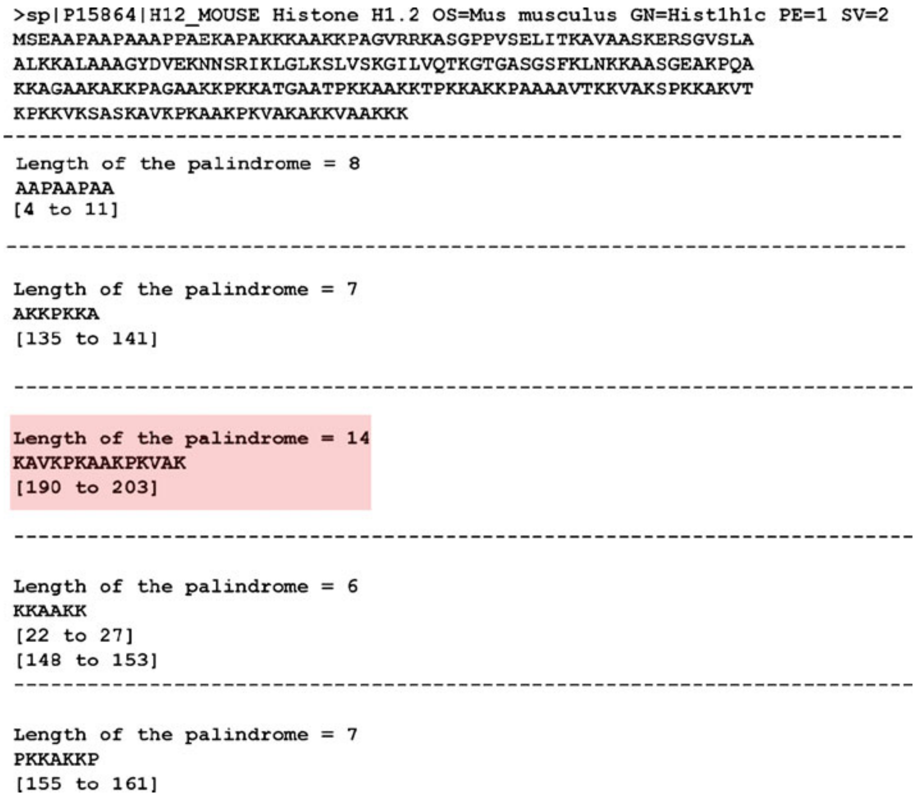


Figure 1. Schematic representation of the proposed algorithm for a sequence ABBSBBAB. Two consecutive pairs $A_{(0)}$ $A_{(6)}$ and $B_{(1)}$ $B_{(5)}$ are first taken into account. Since the end position of the second pair is less than the first, it is ignored and the iteration continues. When the algorithm encounters the sequence 'BAB' ($B_{(5)}A_{(6)}B_{(7)}$), it is considered as an overlapping palindrome because the end position $B_{(7)}$ is greater than the end position of the first pair $A_{(6)}$.

2.2 *Removal of all sub-palindromes*

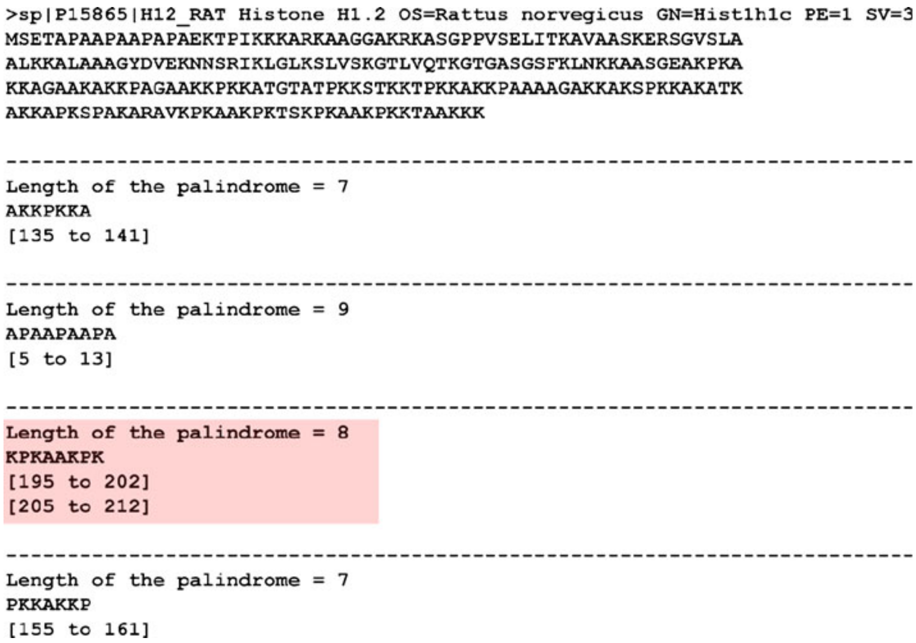
In the second part of the algorithm, the vector is sorted with reference to the start position of the paired palindromic positions. The sub-palindromes are removed by checking the end positions of the consecutive palindromes (res[i].second and res[j].second; j=i+1). Since the vector is sorted with respect to the start position, if res[j].second is less than or equal to res[i].second, then it is a sub-palindrome. Therefore, the sequences with end positions less than or equal to the end positions of their previous palindromic sequences are excluded. This way only the sub-palindromes will be removed leaving behind individual and over lapping palindromes.

For a short sequence ABBSBBAB (figure 1), the positions of the first two consecutive pairs (A(0) A(6) and B(1) B(5)) are taken and their end positions are compared. Since the end position of the second pair is less than the first, it is removed from the vector res. Now, the first pair is compared with the third pair (B(2) and B(4)). Here again, the end position of the third pair lies within the end position of the first pair. Thus, the corresponding end positions are removed. However, the end position of the sequence 'BAB' (B(5)A(6)B(7)) is considered, because it lies after the end position of the sequence ABBSBBA (A(6)). The sequence 'BAB' is an overlapping palindrome. Therefore, the algorithm prints two palindromes 'ABBSBBA' and 'BAB' as output.



```
>sp|P15864|H12_MOUSE Histone H1.2 OS=Mus musculus GN=Hist1h1c PE=1 SV=2
MSEAAPAAPAAAPPAEKAPAKKKAAKKPAGVRRKASGPPVSELITKAVAASKERSGVSLA
ALKKALAAAGYDVEKNNSRIKLGLKSLVSKGILVQTKGTGASGSFKLNKKAASGEAKPQA
KKAGAAKAKKPAGAAKKPKKATGAATPKKAAKKTPKKAKKPAAAAVTKKVAKSPKKAKVT
KPKKVKSASKAVKPKAAKPKVAKAKKVAAKKK
-----------------------------------------------------------------------
Length of the palindrome = 8
AAPAAPAA
[4 to 11]
-----------------------------------------------------------------------

Length of the palindrome = 7
AKKPKKA
[135 to 141]

-----------------------------------------------------------------------

Length of the palindrome = 14
KAVKPKAAKPKVAK
[190 to 203]

-----------------------------------------------------------------------

Length of the palindrome = 6
KKAAKK
[22 to 27]
[148 to 153]
-----------------------------------------------------------------------

Length of the palindrome = 7
PKKAKKP
[155 to 161]
```

Figure 2. A total of five palindromic sequences of varying lengths 8, 7, 14, 6 and 7 have been identified by the algorithm in the mouse histone H1. The palindromic sequence coloured in pink is one of the longest palindromic sequences reported so far in the literature.



```
>sp|P15865|H12_RAT Histone H1.2 OS=Rattus norvegicus GN=Hist1h1c PE=1 SV=3
MSETAPAAPAAPAPAEKTPIKKKARKAAGGAKRKASGPPVSELITKAVAASKERSGVSLA
ALKKALAAAGYDVEKNNSRIKLGLKSLVSKGTLVQTKGTGASGSFKLNKKAASGEAKPKA
KKAGAAKAKKPAGAAKKPKKATGTATPKKSTKKTPKKAKKPAAAAGAKKAKSPKKAKATK
AKKAPKSPAKARAVKPKAAKPKTSKPKAAKPKKTAAKKK

-----------------------------------------------------------------------
Length of the palindrome = 7
AKKPKKA
[135 to 141]

-----------------------------------------------------------------------
Length of the palindrome = 9
APAAPAAPA
[5 to 13]

-----------------------------------------------------------------------
Length of the palindrome = 8
KPKAAKPK
[195 to 202]
[205 to 212]

-----------------------------------------------------------------------
Length of the palindrome = 7
PKKAKKP
[155 to 161]
```

Figure 3. The palindromic sequence highlighted in pink occurs twice in the protein separated by two amino acids T and S. This is an example for tandemly repeated palindromic sequences.

3. Case studies

Histones are DNA-binding proteins involved in the packaging of DNA strands into highly condensed structures (Cox and Lehninger 2005). In addition, most parts of the amino acid sequence (minimum of 50%) have been identified to form palindromes (Ohno 1992). Mouse Histone H1 (212 amino acid residues) protein contains one of the longest palindromic sequences reported so far. Hence, this protein sequence was given as input to the proposed algorithm and the minimum number of residues in the palindrome was set to be 5. A total of 5 palindromic sequences (figure 2) of lengths 8, 7, 14, 6 and 7 were displayed in the output file. As mentioned above, the sequence ‘KAVKPKAAKPKVAK’, a 14-residues palindrome (figure 2) and one of the longest palindromic peptide sequences, occurs (Ohno 1992) in the low-complexity region. The presence of the palindrome in the low-complexity region supports the results reported by Sheari *et al*. (2008). However, a quick scan of the protein sequences available in the Protein Databank (PDB) resulted in palindromes of longer lengths; a maximum length of 31 amino acid residues was observed in PDB-ids: 2DRT and 2G66 (results not shown).

Interesting results were obtained when a rat histone H1 was given as the input sequence. Out of the 4 palindromic peptide sequences identified by the algorithm (figure 3) (minimum number of residues in the palindrome was set as 5), the octapeptide palindrome sequence is unique, because it is present twice (adjacent to each other) in the same sequence and it is separated by only two amino acid residues T and S (KPKAAKPK **TS** KPKAAKPK). The occurrence of such tandemly repeated palindromes may be perceived as the outcome of mutation events (addition, deletion or substitution) which would result in the separation of a single long palindromic sequence into two smaller palindromes (Ohno 1992).

To the proposed algorithm, the parallel β-helix-repeat-containing protein (36,805 amino acid residues) from *Chlorobium chlorochromatii* CaD3 was given as the input sequence and the minimum number of residues in the palindrome was given as 2. The algorithm was able to identify 444 palindromic sequences in 3.6 s (results not shown). This shows that the proposed algorithm is very fast and does not impose any limit on the number of characters (number of amino acid residues) in the input sequence.

4. Conclusion

Many algorithms are available for the identification of palindromic nucleotide sequences; however, there is no algorithm available for the identification of palindromic peptide sequences. Thus, in view of the above, an algorithm to find palindromic sequences in a single protein sequence has been developed. This would help biologists tackle many unanswered questions about palindromic peptide sequences available in various databases. The source code of this algorithm may be obtained from KS on request.

Acknowledgements

The authors gratefully acknowledge the facilities offered by the Bioinformatics Centre, Supercomputer Education and Research Centre and the Interactive graphics facility. One of the authors (KS) thanks the Department of Information Technology (DIT) for the financial support in the form of a research grant.

References

Cheng GH, Nandy A, Clerk S and Skoultchi AI 1989 Different 3′-end processing produces two independently regulated mRNAs from a single H1 histone gene. *Proc. Natl. Acad. Sci. USA* **86** 7002–7006

Cox MN and Lehninger DR 2005 *Lehninger principles of biochemistry* (San Francisco: W.H. Freeman)

Engberg J and Klenow H 1977 Palindromic arrangement of specific genes in lower eukaryotes. *Trends Biochem. Sci.* **2** 183–185

Engberg J, Andersson P, Leick V and Collins J 1976 Free ribosomal DNA molecules from *Tetrahymena pyriformis* GL are giant palindromes. *J. Mol. Biol.* **104** 455–470

Giel-Pietraszuk M, Hoffmann M, Dolecka S, Rychlewski J and Barciszewski J 2003 Palindromes in proteins. *J. Protein Chem.* **22** 109–113

Hoffmann M and Rychlewski J 1999 Searching for palindromic sequences in primary structure of proteins. *Comput. Methods Sci. Tech.* **5** 21–24

Jaseja M, Mergen L, Gillette K, Forbes K, Sehgal I and Copié V 2005 Structure- function studies of the functional and binding epitope of the human 37 kDa laminin receptor precursor protein. *J. Peptide Res.* **66** 9–18

Karrer KM and Gall J 1976 The macronuclear ribosomal DNA of *Tetrahymena pyriformis* is a palindrome. *J. Mol. Biol.* **104** 421–453

Kazim AL 1993 Identification of putative internalization signals in prion proteins. *FEBS Lett.* **331** 1–3

Lacroix E, Viguera AR and Serrano L 1998 Reading protein sequences backward. *Fold Des.* **3** 79–85

Mittl PR, Deillon C, Sargent D, Liu N, Klauser S, Thomas RM, Gutte B and Grutter MG 2000 The retro-GCN4 leucine zipper sequence forms a stable three-dimensional structure. *Proc. Natl. Acad. Sci. USA* **97** 2562–2566

Ohno S 1990 Intrinsic evolution of proteins: The role of peptidic palindromes. *Rivista di Biologia - Biology Forum* **83** 287–291

Ohno S 1992 Of palindromes and peptides. *Hum. Genet.* **90** 342–345

Ohno S 1993 A song in praise of peptide palindromes. *Leukemia* **7** S157–S159

Pal-Bhowmick I, Pandey RP, Jarori GK, Kar S and Sahal D 2007 Structural and functional studies on Ribonuclease S, retro S and retroinverso S peptides. *Biochem. Biophys. Res. Commun.* **364** 608–613

Pan PK, Zheng ZF, Lyu PC and Huang PC 1999 Why reversing the sequence of the α domain of human metallothionein-2 does not change its metal-binding and folding characteristics. *Eur. J. Biochem.* **266** 33–39

Rai J 2007 Retroinverso mimetics of S peptide. *Chem. Biol. Drug Des.* **70** 552–556

Roulland-Dussoix D and Boyer HW 1969 The *Escherichia coli* B restriction endonuclease. *Biochim. Biophys. Acta* **195** 219–229

Sheari A, Kargar M, Katanforoush A, Arab S, Sadeghi M, Pezeshk H, Eslahchi C and Marashi SA 2008 A tale of two symmetrical tails: structural and functional characteristics of palindromes in proteins. *BMC Bioinformatics* **9** 274

Shukla A, Raje M and Guptasarma P 2003 A backbone-reversed form of an all-β α- crystallin domain from a small heat-shock protein (retro-HSP12.6) folds and assembles into structured multimers. *J. Biol. Chem.* **278** 26505–26510

Sulkowski E 1992 Aromatic palindrome motif in prion proteins. *FASEB J.* **6** 2363

Suzuki M 1992 DNA-bridging by a palindromic α-helix. *Proc. Natl. Acad. Sci. USA* **89** 8726–8730

Vogt VM and Braun R 1976 Structure of ribosomal DNA in *Physarum polycephalum*. *J. Mol. Biol.* **106** 567–587

MS received 22 August 2012; accepted 04 January 2013

Corresponding editor: Reiner A Veitia