The MSU Research Consultancy Project Proposal

Applying Signal Processing based Algorithms for Recognizing Short Tandem Repeat Regions in DNA Sequences for diseases like Alzheimer's or Cancer

by

Dr. Mamta Chandraprakash Padole

Associate Professor

Department of Computer Science and Engineering

Faculty of Technology and Engineering

The Maharaja Sayajirao University of Baroda

29th December, 2016

Applying Signal Processing based Algorithms for Recognizing Short Tandem Repeat Regions in DNA Sequences for diseases like Alzheimer's or Cancer

Keywords:

- DNA Sequences
- Short Tandem Repeats
- Signal Processing
- Wavelet Transforms
- Haar Wavelet Transforms

DNA

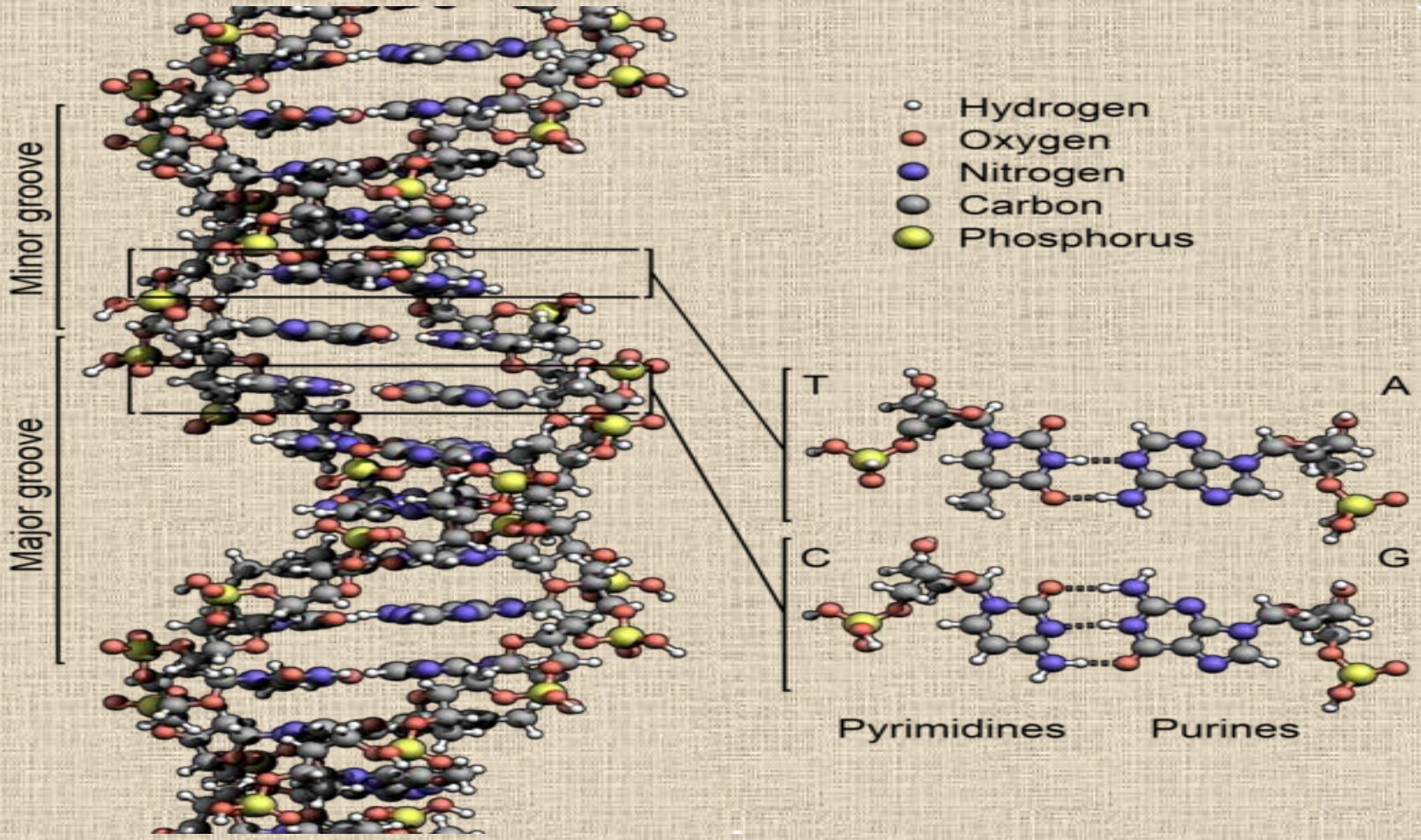


Fig. DNA, as a chain of smaller molecules or nucleotide bases
Source: http://upload.wikimedia.org

DNA Sequence

- >gi|194719540|ref|NC_007331.3|NC_007331 BosTaurus chromosome X, reference assembly (based on Btau_4.0), whole genome shotgun sequence

ATTCTCCAGGCCAGAATACTGGGATGGGTAGCCTTTCCCTCCTCCAGG
TGATCTTCCCAACCCAGGAATCGAACCCAGGTCTCCTGCATTGCAGGT
AGATTCGTTACCAGCTGAGACACAAGGGAAGCCCAAGAGTATTGGAGT
GGGTAGCCTATCCCTTCTCCAGCGGATTTTCCCAATCCAGGAATCAAA
CTGGGGTCTCCTGCATTGCAGGTGGATTCTTTACCAACTGAGCCACAA
GGGAAGCCCCTCACAAATATTGGTCCTATGAAAATCAGCTCCCTTGCA
GTTACAGAAAGGAGCAAATTGTATATAAATTTCCTCAAAAATTTCCCATT
CCTGGAGCATTCCTGTTTCTAGAGCTTTGACACTGTTTGACCTGTTTCA
TAGCTCCCTAGAAAAATCCCACTCCAAACCTGTTATTATTAGAATCTGA
CTCAGCACTTGCTTTGCAAAACAGCCCTGTCCCCAGGTCATTTGTCAAA
AACAATCAGTGGCAATTGTTTAACACCTTAGTTGCTTGAGGCAGCAATA
ACAGTTGGAGCCAAAAAGAGGCTAACCAAGAAGCTAAAAAAAAACAAT
GATGTGGGGGGAAAAAAGACAAATATAGGGGAATTTGAAAAGCTCTGA
G

DNA

- DNA sequence provides blueprint for inheritance.
  - A basic physical unit that is arranged sequentially in form of genes.
  - Genes are passed from parents to offspring with information necessary to specify traits.
  - Genes are arranged on structures called chromosomes.
- The De-oxy Ribo-Nucleic Acid (DNA) is the chemical name for the molecule that carries genetic information in all living beings.
- Double-helix structure of nucleotides, held together by chemical bonds between the nucleotides.
- Each sugar molecule is attached to one of four bases
  - Adenine (A) Cytosine (C)
  - Guanine (G) Thymine (T).
- The sequence and number of bases is what creates diversity.
- DNA is transcribed into RNA and then translated into proteins.

Short Tandem Repeats (STRs)

- Short Tandem Repeats (STRs)
  - Microsatellites
  - Short Sequence Repeats (SSRs)
  - Variable number Tandem Repeats (VNTR)
- STRs are contiguously placed or ubiquitously distributed, multiple and approximate copies of pattern of nucleotides, in DNA sequences
- They are nucleotide sequences in DNA of 1–6 bp unit length, distributed randomly in eukaryotic and prokaryotic genomes and are highly polymorphic.
- Eg:
- GTACTATGTA**TTTTTTTTTTTTTTTTTTTT**ACGA**GTGTGTGT**CAT
- GTA**TCATCATCATCATCATCATCATCA**CA**TTTT**CA**GTACGTACG TAC**TATGTA

Sample of STRs

Generic Term/ Biological Term for Type of Repeat	***Length of Repeat***	***Repeat Sequence***	***Annotation Repeat Unit and Its Frequency)***
Homopolymeric or Monomer/ Perfect	9	5'-ACGA**TTTTTTTTT**CA-3'	5'-(T)9-3'
Multimeric, Dimer/ Perfect	6	5'-ACC**ATATATATATAT**GA-3'	5'-(AT)6-3'
Multimeric,Trimer/ Perfect	5	5'-**CAGCAGCAGCAGCAG**-3'	5'-(CAG)5-3'
Multimeric,Tetramer/ Perfect	4	5'-**GATCGATCGATCGATC** -3'	5'-(GATC)4-3'
Multimeric,Pentamer/ Perfect	4	5'-**ATGCCATGCCATGCCATG CC** -3'	5'-(ATGCC)4-3'
Multimeric, Imperfect Heterogeneous	Variant	5'-*GCC GCC GCC* **GATC GATC** AT AT AT AT -3'	5'-(GCC)3 (GATC)2 (AT)4-3'
Multimeric, Heterogeneous/ Imperfect and Interrupt	Variant	5'- ***GCC GCC GCC*** T **GATC GATC** GC **AT AT AT AT** -3'	5'-(GCC)3 T (GATC)2 GC (AT)4 -3'

Purpose To Identify STRs

STRs have an impact in:

- Role in Regulation of gene expression
- Mutational dynamics of STRs play a role in human genetic disorders
- Cause several human diseases like *Cancer,* Diabetes, HIV, myotonic dystrophy, spinal and bulbar muscular atrophy, *Friedreich's* ataxia, *Huntington's* disease etc.
- ~ 25% to 30% (3.1 billion base-pairs (3×10^9)) of human genome comprises of repeats
- Population genetic analysis – Sickle Cell disease in South Guj.
- Genetic mapping
- Phylogenetics – Homo Nadali, Homo Sapiens Sapiens
- DNA forensics - Crime Detection

Pattern Mining

- "Pattern mining" is a data mining method that involves finding existing patterns in data.
- *Given a database D with transactions T*1 *. . . TN, determine all patterns P that are present in at least a fraction s of the transactions*
- In this context *patterns* often means association rules.
- The **patterns** generally have the form of
  - Sequences (Linear)
  - Tree structures
- In Genomic DNA sequences, the pattern have a linear form

Signal Processing

- “Signal” is a formal representation of phenomenon evolving over time or space
- “Signal Processing” : Operating on a signal using some function, to extract out the information preserved in the signal.
- Signal processing deals with identifying some function which can be applied
  - Representation
  - Transformation
  - Manipulation

  of a signal and its contained information.
- “Transform of a signal” is just a different form or a method of representing the signal, without altering its meaning.
- Represented as infinite sum of scaled and shifted unit impulses

$$\int_{t=-\infty}^{\infty} x(t).dt = \lim_{\delta \to 0} \sum_{k=-\infty}^{\infty} x(k.\delta).\delta$$

- Shannon’s Representation (Time Domain): Time vs. Amplitude
- Fourier’s Representation (Frequency Domain): Frequency vs. Amplitude
- Wavelet: Time vs. Scale (Scale is an inverse of Frequency)

Wavelet Transforms (WT)

- **Linear** transformation, Useful in Analyzing the **non-stationary** signals, It can be applied **for lossless** transformations
- Represents the signal in Time-Scale Domain **(Scale is an inverse of frequency)**
- **Time or Positional Information is preserved** unlike in Fourier Transforms, **Frequency** content of a signal can be acquired using WT as it **preserves both time & frequency** contents
- It transforms a signal or data, into co-efficients, on a basis of wavelet functions.
- Wavelet Transform of a signal X can be represented as

  $W_T = X.\ W$ where, $W = [\Phi(t); \psi(t)]$
- ***Thus, Wavelet Transform is the Convolution of a Signal, with the Wavelet Function or Mother Wavelet***
- Applications of Wavelet Transforms are **.jpeg, .mpeg,** reduces **transmission time** in mobile apps, ElectroCardiogram analysis

Haar Wavelet Transforms

- The ith element in a Ca and Cd vector after decomposition level j, can be obtained as:
  - $C_{aj} = 1/\sqrt{2}\ (X_j(2i-1)) + 1/\sqrt{2}\ (X_j(2i))$
  - $C_{dj} = 1/\sqrt{2}\ (X_j(2i-1)) - 1/\sqrt{2}\ (X_j(2i))$
- In Haar Wavelets, the length of original signal is expected to be of the power of 2.
- Length of the transformed vector containing the detailed co-efficient Cd, is usually n/2j, where j is the decomposition level.
- The decomposition using Haar wavelets can be performed until the resolution (number of approximation co-efficient) becomes one or resolution level zero. Number of detailed co-efficients at each level j is equal to n/2j .
- As per Nyquist's rule, With every transform, and by performing down-sampling, half the values (keep the even element) of the given signal can be discarded. This discarding reduces the length i.e. number of elements into half, on each transformation.
- **Hence, optimizing the search, with time complexity of *O(log n).***

Haar Wavelet Transform

Transformation Level or Decomposition Level (j)	Scale $a = 2^j$	Resolution r = 1/a)	Length of Signal (L)	Approximation Level (A_j),	Averages / Approximate Co-efficient (C_a)	Detail Co-efficient Level (D_j)	Differences / Detail Co-efficient (C_d)
(a)	(b)	(c)	(c)	(d)	(e)	(f)	(g)
Original Signal Level 0	$2^0 = 1$	$1/2^0 = 1$	16	A_0	[18,16,7,2,6,6,6,6,4,14,3,14,12,20,12,20]	D_0	-
1	$2^1 = 2$	$1/2^1 = ½$	8	A_1	[34/√2, 9/√2, 12/√2, 12/√2, 18/√2, 17/√2, 32/√2, 32/√2]	D_1	[2/√2, 5/√2, 0,0, -10/√2, -11/√2, -8/√2, -8/√2]
2	$2^2 = 4$	$1/2^2 = 1/4$	4	A_2	[43/2, 12, 35/2, 32]	D_2	[25/√2, 0, 1/√2, 0]
3	$2^3 = 8$	$1/2^3 = 1/8$	2	A_3	[67/2√2, 99/2√2]	D_3	[19⁄2√2, -29⁄2√2]
4	$2^4 = 16$	$1/2^4 = 1/16$	1	A_4	[166/4]	D_3	[-32⁄4]

Literature Review - STRs

- The *in-vitro* approach or wet-lab methods involve expensive probe hybridization.
- The *in-silico* approach or computational methods include study of regional distribution bias or putative association with genomic features.
- These *in-silico* investigations are widely used instead of expensive *in-vitro* method .
- Existing tools for detecting short tandem repeats are:
  - *MISA, REPuter, Sputnik, RepeatMasker*
- Existing tools primarily use String-comparison based algorithms.
  - Regular-expression, Hamming distance, Dynamic Programming
  - k-mer with suffix trees and k-tuples, Seed extension technique
- String-based and other approaches need input parameters as:
  - Pattern, Pattern size, Reference sequence, Exponential Complexity
- These algorithms are
  - computationally complex
  - memory intensive

Proposed Algorithm Identifying STR Regions

1. Read the Fasta File
2. Convert each sequence into Numerical Representation
3. Perform 1st level of Haar Wavelet Transform of each Numerical Representation of the sequence
4. Identify the series of zeros in Detailed Co-efficient of 1st level of Transform
5. Identify the start and end position of zeros and hence the length of zeros.
6. Multiply the zero position p by $(2^i) - 2i - 1)$ to find the original starting position of the repeat. First level of transform will generate information about repeat regions of monomers in the given sequence
7. Perform Steps 3 to 6 for further decomposition levels to recognize di-mer, tri-mers and tetramers in the sequence

Dipole Moments value for Nucleotide Bases

- A - 0.4629
- G - 6.488
- C - 3.943
- T - 1.052
- The use of dipole-moment property which is a single indicator for nucleotide base
- It reduces the computational overhead by 75% compared to the conventional two-base or four-base binary sequence representation of nucleotide sequence.
- Only numerical representations can be applied for transformations. (Various Encoding Schemes: Single Galois Indicator, Electron-Ion Interaction Pseudo Potential (EIIP), Molecular Mass, Atomic Number etc.)

Using WT to Identify STRs

Consider a DNA sequence as follows:

ACGATATTTTTTTTTTCAGATGACACACACACCTAGGCT

Numerical Representation Using Dipole Moments Is:

0.462900000000000	3.94300000000000	6.48800000000000
0.462900000000000	1.05200000000000	0.462900000000000
1.05200000000000	1.05200000000000	1.05200000000000
1.05200000000000	1.05200000000000	1.05200000000000
1.05200000000000	1.05200000000000	1.05200000000000
1.05200000000000	3.94300000000000	0.462900000000000
6.48800000000000	0.462900000000000	1.05200000000000
6.48800000000000	0.462900000000000	3.94300000000000
0.462900000000000	3.94300000000000	0.462900000000000
3.94300000000000	0.462900000000000	3.94300000000000
0.462900000000000	3.94300000000000	3.94300000000000
1.05200000000000	0.462900000000000	6.48800000000000
6.48800000000000	3.94300000000000	1.05200000000000

Using WT to Identify STRs

ACGATATTTTTTTTTTTCAGATGACACACACACCTAGGCT

1st Level Decomposition – Detail Coefficients

-2.46080230920730 4.26038906732707

-1.79958675811976 0 0 0 0 0

2.46080230920730 4.26038906732707 -3.84383246253007 -2.46080230920730-2.46080230920730 -2.46080230920730

-2.46080230920730 -2.46080230920730

2.04424570441031 -4.26038906732707

1.79958675811976 0.416556604796995

2nd Level Decomposition – Detail Coefficients

-1.27250000000000 4.16350000000000

0 0 -1.27250000000000 1.56705000000000

0 0 -0.977950000000000 4.45805000000000

Using WT to Identify STRs

ACGATATTTTTTTTTTTCAGATGACACACACACCTAGGCT

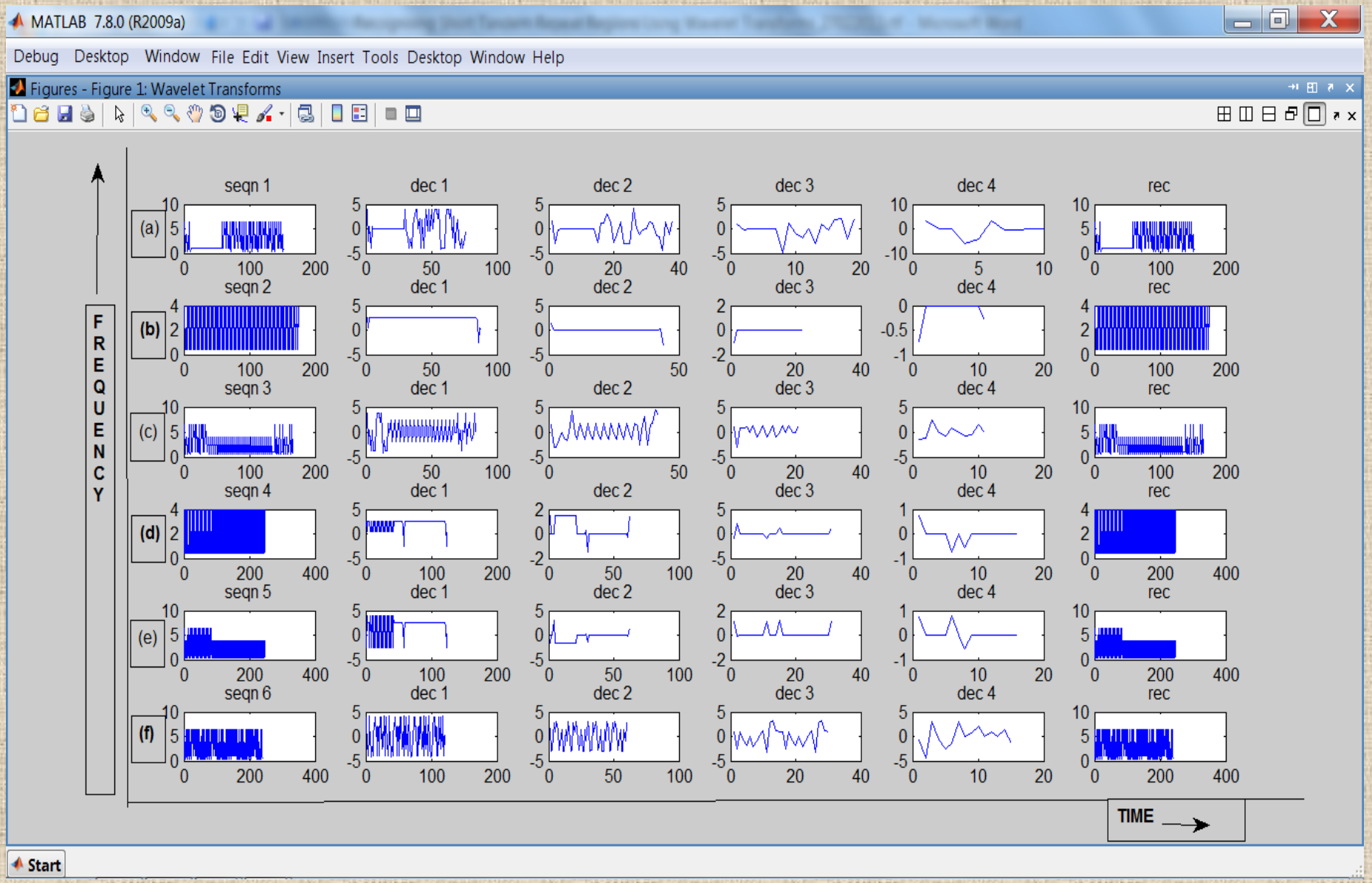
Output in a file:

Total Processing time : 3.099401e-001
Wavelet Analysis time : 4.663242e-002
Sequence FYRUZ7J01B3YZA contains STR T of size 10 at Start Position 7 and End Position 16
Sequence FYRUZ7J01B3YZA contains STR C of size 2 at Start Position 32 and End Position 33
Sequence FYRUZ7J01B3YZA contains STR G of size 2 at Start Position 36 and End Position 37
Sequence FYRUZ7J01B3YZA contains STR AC of size 5 at Start Position 23 and End Position 32
File Writing time 1.130018e-003 sec

Sample of Input Data

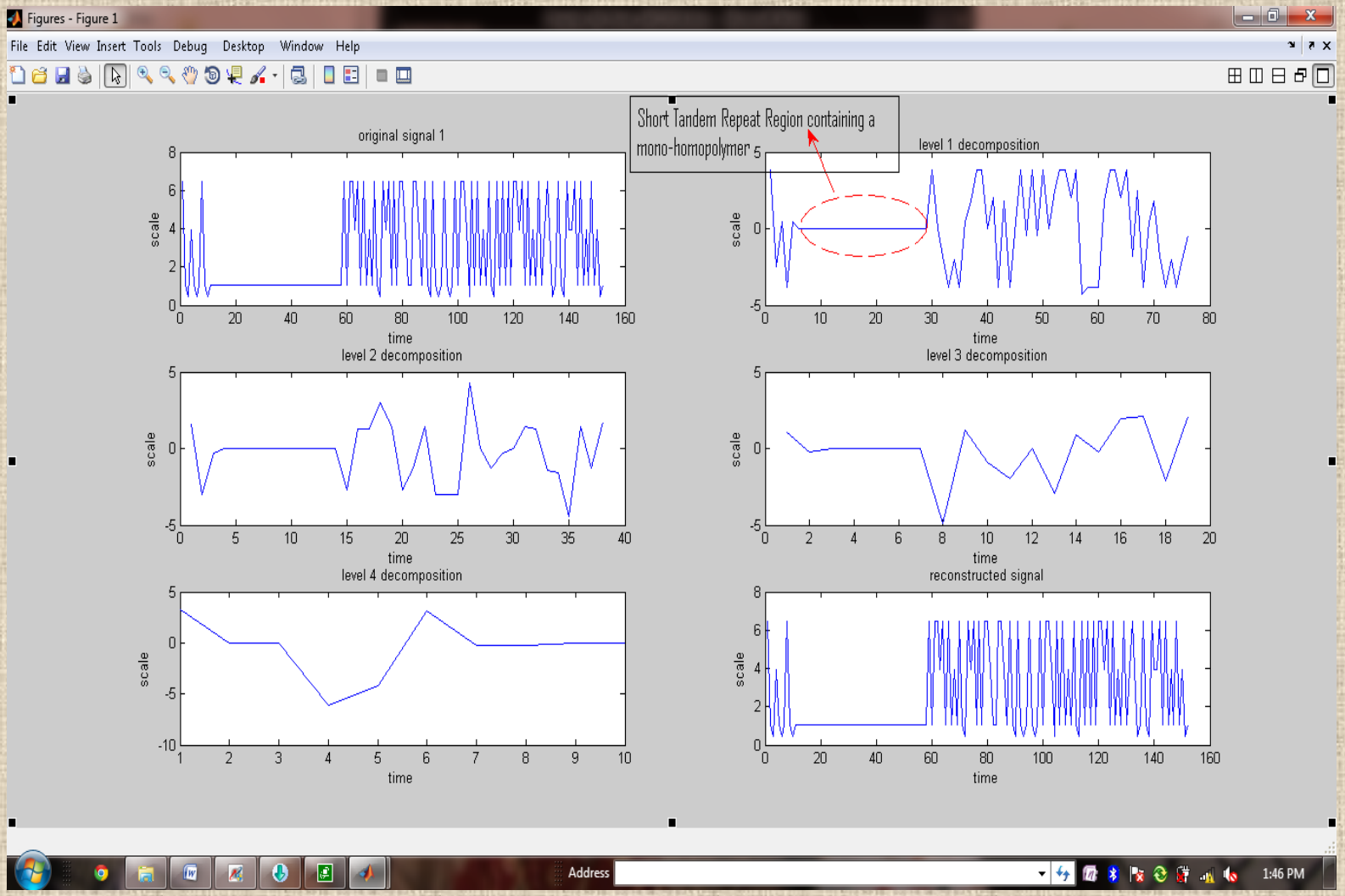
- >FYRUZ7J01B3YZA rank=0001782 x=748.0 y=836.5 length=152
- GTACTATGTA**TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT**GTGGCGTGTCTGTAGCGTGTGGCTTGGC TGTAGTATGTATGTGGCAGTGTCTGTAGTGTGTGGCGTGTCTGTCGCATGTAGCCGTGTCTGTCAT
-
- >FYRUZ7J01A1PTO rank=0003965 x=312.0 y=1898.5 length=175
- CAT**ACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACA CACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACA CACACACACACAC**TACTAC
-
- >FYRUZ7J01A1PT1 rank=0003965 x=399.0 y=2198.5 length=167
- GTACTATGTATGTGGCGTGTCTGTAGCGTGTGGCT**TCATCATCATCATCATCATCATCATCATCATCATCATCATCATCATC ATCATCATCATCATCATCATCATCATCATCATCATCATCATCATCATCATCA**TTTTGTACTATGTATTGTACTactatcgATGTA
-
- >FYRUZ7J01A1PT2 rank=0003965 x=442.0 y=2898.5 length=248
- CATA**CACACACA**TA**TACATACATACATACATACATACATACATACATACATACATACATACATACATACATACATACATAC ATACA**CACACACACACACACACACACACACA**CTACTA**CACACACACACACACACACACACACACACACACACACACACACA CACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACC TACTA
-
- >FYRUZ7J01A1PT3 rank=0003965 x=467.0 y=3908.5 length=248
- ACGTCACACAACGTTA*ACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTA CGT***CACACACACACACACACACACACACACACTACTACACACACACACACACACACACACACACACACACACACACACA CACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACA CACA**CTACTA
-
- >FYRUZ7J01A1PT4 rank=0003965 x=490.0 y=4208.5 length=240
- GTACTATGTAGACGTCAGTCTGTAGCGTGTGGCTTGGCTGTAGTATGTactatcg**ATGTAGTACTATGTAGACGTCAGTCTGT AGCGTGTGGCTTGGCTGTAGTATGTactatcg**ATGTAGTACTATGTAGACGTCAGTCTGTAGCGTGTGGCTTGGCTGTAGTAT GTactatcg**ATGTAGTACTATGTAGACGTCAGTCTGTAGCGTGTGGCTTGGCTGTAGTATGTactatcg**ATGTA

Graphical Output – Short Tandem Repeat Regions



Mamta C. Padole

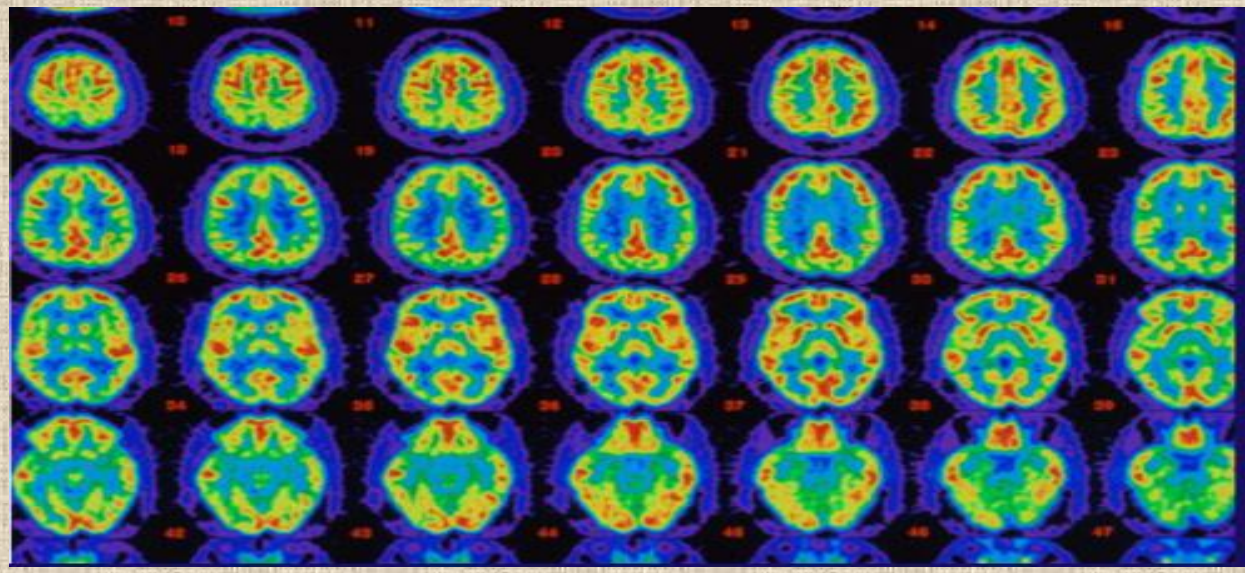
Graphical Output – Short Tandem Repeat Regions



Development Tools & Languages Used

- Matlab R2010b Software / Scilab
  - Wavelet Toolbox
  - BioInformatics Toolbox
- Java Programming
- MySQL Database

Alzheimer's



Advances in brain imaging

Advanced imaging is opening a new window to the brain. PET scans with radiotracers such as Pittsburgh Compound B — developed with Alzheimer's Association support — can show how key pathologies accumulate over time, paving the way to diagnose Alzheimer's earlier and monitor treatment and progression.

Boosting brain cell communications

The brain's power lies in its synapses — 100 trillion connections where nerve cells pulse information to one another. Current Alzheimer treatments temporarily support these vital cell-to-cell signals. But they don't prevent cell decline and death. The goal of new treatments is to keep cells alive and thriving.

Facts about Alzheimer's

- Disease that is 6th Leading Cause of Death according to US studies [http://www.alz.org/facts/overview.asp]
- 1 in 3 Senior citizens with Alzheimer's Dementia
- > 5 million Americans are living with Alzheimer's
- Every 66 seconds, someone in US develops the disease
- It kills more people than Breast and Prostrate Cancer combined
- In 2016 alone approx $236 Billion Estimated cost to the Nation for Alzheimer's Health Care
- Approx $5000 / year is spent by a family to take care of Alzheimer's patient at home
- In 2015, More than 15 million people provided UnPaid Care
- Estimated to 18.1 Billion Hours of UnPaid Care

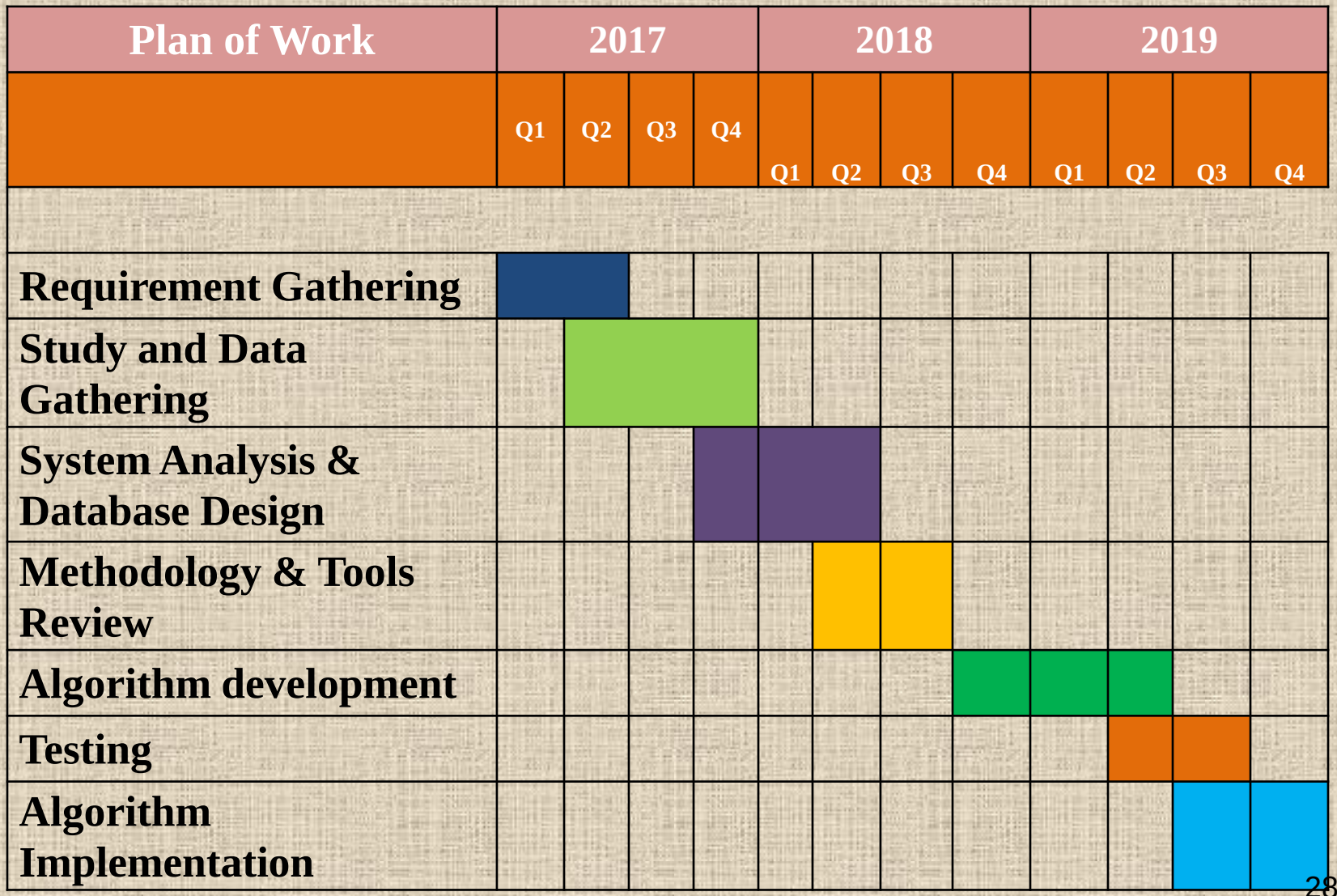
National / International Status

- Research at National/International level to combat diseases like Alzheimer, Cancer
- Research is of global relevance, is not confined to any region or locality
- Besides with development of Genetics, and Genomic sequencing, the need to develop optimized methods to detect these diseases have arisen.
- International Research organizations like NCBI, EMBL, Plos,
- Stressful Life, Social transformations, Unhealthy Life Style
- This, causes lot of genetic mutation/exchange and hence, could be one of the reasons for causing the Genetic diseases like Cancer or Alzheimer (Chromosome 21, 14, 1) have several repeat regions causing genetic mutation affecting Brain Cells.

Methodology

- The study of diseases and collecting relevant information and data, Performing Analysis of the system
- The data analysis and database will be designed to store the content
- Methodology and tools will be reviewed for specific diseases
- Develop and redesign Signal Processing based algorithms for identifying disease causing Short Tandem Repeat Regions in DNA sequences.
- The testing of the algorithm using several sample data will be done.
- The algorithms will be applied to actual DNA sequences.

Plan of Work


Plan of Work
2017
2018
2019
Q1
Q2
Q3
Q4
Q1
Q2
Q3
Q4
Q1
Q2
Q3
Q4
Requirement Gathering
Study and Data Gathering
System Analysis & Database Design
Methodology & Tools Review
Algorithm development
Testing
Algorithm Implementation

Year-Wise Expenditure Estimate

Sr . No.	Items	2016-2017	2017-2018	2018-2019	Total
1	Books / Articles	5,000	5,000	2,000	**12,000**
2	Contingencies	3,000	10,000	5,000	**18,000**
3	Consumables	5,000	5,000	3,000	**13,000**
4	Field Work / Travel (for data collection)	12,000	15,000	5,000	**32,000**
5	Minor Instruments	5,000	15,000	5,000	**25,000**
	Total	**30,000**	**50,000**	**20,000**	**1,00,000**

Conclusion

- Designing and development of algorithms using an altogether innovative approach of Signal Processing using Wavelet Transforms as applied for Genetics and Health Care Applications

- Applying Signal Processing to the newly emerging domain of BioInformatics, with an urge for inter-disciplinary research work

- To support and improvise existing methods in various domains by optimizing processing time with minimum resources.

List of Publications

- **Recognizing Short Tandem Repeat Regions In Genomic Sequences Using Wavelet Transforms**
  - Published by IEEE, DOI 10.1109/MCSI.2014.50, Pg. 288-294, ISBN: 978-1-4799-4744-7
  - *Indexed **IEEE Xplore***
  - *Indexed by **ACM Digital Library**: http://dl.acm.org/citation.cfm?id=2763062*
  - *Indexed by **Google Scholar***
- **Recognizing Artificial Duplicate Reads in 454 Pyrosequencing Using Wavelet Transforms**
  - International Journal of Advanced Computing (ISSN: 2051-0845), Recent Science Public.
  - IMPACT FACTOR: 2.31
- **Signal Processing Approach for Recognizing Identical Reads From DNA Sequencing of Bacillus Strains**
  - International Journal of Computer Engineering (IOSR-JCE) (e-ISSN: 2278-0661, p- ISSN: 2278-8727)
  - Paper Indexed by: CrossRef (DOI 10.9790/0661-01011924), ANED, ESCI, Google Scholar
- **Distributed Computing for Structured Storage, Retrieval and Processing of DNA Sequencing Data**
  - International Journal of Internet and Web Technology (ISSN: 2051-6878)
  - IMPACT FACTOR: 1.89
- **Dimensionality Reduction of DNA Sequences Using Wavelet Transforms**
  - Pg. 145-152 of Conference Proceedings in Recent Advances in Computer Engineering Series - 18, ISSN:1790-5109,ISBN:978-960-474-354-4
  - http://www.wseas.us/e-library/conferences/2013/Nanjing/ACCIS-23.pdf

References

- Alan v. Oppenheim, Ronald W. Schafer, John R. Buck, Discrete-Time Signal Processing, Prentice Hall, 2nd Edition, 1998
- http://www.mathworks.com [Matlab documentation].
- Ingrid Daubechies, “Ten Lectures On Wavelets”, 1992
- J.K. Meher, M. R. Panigrahi, G. N. Dash, P. K. Meher, “Wavelet Based Lossless DNA Sequence Compression For Faster Detection Of Eukaryotic Protein Coding Regions” I.J. Image, Graphics & Signal Processing 2012, 47-53
- Leland Wilkinson, Statistics and Computing - The Grammar of Graphics, Springer, 28-Jan-2006
- Lecture Notes: Professor David Heeger, Signals, Linear Systems, and Convolution, September 26, 2000
- http://www.r2labs.org/references/Convolution.pdf
- Lecture Notes : Willard Miller, Introduction to the Mathematics of Wavelets, May 3, 2006
- Robi Polikar, The Wavelet Tutorial, Rowan University, 2001.
- Amara Graps, An Introduction to Wavelets

References

- G. H. Watson, ‘Application of Mathematical Signal Processing Techniques to Mission Systems”, Paper presented at the RTO SC1 Lecture Series on held in Kiiln, Germany, I-2 November 1999 (Introduction to Wavelet Analysis)
- D. Lee Fugal, Conceptual Wavelets in Digital Signal Processing, Space & Signals Technologies LLC, 2009 (www.ConceptualWavelets .com)
- Charu C. Aggarwal, “On The Use Of Wavelet Decomposition For String Classification,” Springer - Data Mining And Knowledge Discovery, 10, 117–139, 2005
- Web Content: Musawir Ali, An Introduction to Wavelets and the Haar Transform (www.cs.ucf.edu/~mali/haar/)
- http://courses.ae.utexas.edu/ase463q/design_pages/fall02/wavelet/4_wavelet_theory .htm
- Moharir P.S., “Pattern recognition transforms,”New York: Wiley, 1992
- Christopher Moretti, Michael Olson, Scott Emrich, and Douglas Thain, Highly Scalable Genome Assembly on Campus Grids, Portland, Oregon, USA, MTAGS ’09 2009 ACM 978-1-60558-714-1/09/11
- P. Prandoni , M. Vetterli, Signal Processing For Communication, EPFL Press, 2008
- Charu C. Aggarwal, J. Han, “Frequency Pattern Mining”, Springer Publication

References

- Benson G., “Tandem repeats finder: a program to analyze DNA sequences,” PMC, PubMed Nucleic Acids Research, Vol. 27. No. 2 1999; 27:573–80
- Li YC, Korol AB, Fahima T, “Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review,” Mol Ecol, 2002; 11:2453–65
- Angelika Merkel, Neil Gemmell, “Detecting short tandem repeats from genome data: opening the software black box,” Briefings in Bioinformatics Advance Access, July 10, 2008
- Goldstein DB, Schlotterer C ,“Microsatellites: Evolution and Applications,” Oxford University Press, 1999
- Pearson CE, Edamura KN, Cleary JD, “Repeat instability: mechanisms of dynamic mutations,” Nat Rev Genet, 2005; 6: 729–42.
- Kashi Y, King DG,“Simple sequence repeats as advantageous mutators in evolution,” Trends in Genetics, 2006;22: 253–9
- Moxon ER, Wills C., “DNA microsatellites: agents of evolution? ” Sci Am, 1999; 280:94–9
- Jeffreys, A. J., V. Wilson, and S. L. Thein, “Hypervariable “minisatellite” regions in human DNA,” Nature, 1985 314:67–73
- Nakamura, Y., M. Leppert, P. O’Connell, R. Wolff, T. Holm, M. Culver, C. Martin, E. Fujimoto, M. Hoff, E. Kumlin, and R. White, “ Variable number of tandem repeat (VNTR) markers for human gene mapping ,” Science, 1987 235:1616–1622
- Thiel T, Michalek W, Varshney RK, “Exploiting EST databases for the development and characterization of gene derived SSR-markers in barley (Hordeum vulgare L.),” Theory of Applied Genet, 2003; 106:411–22.

References

- Smit AFA, Green P, “RepeatMasker,”Available at: http://www.repeatmasker.org, 1996
- Goffeau A, Barrell BG, Bussey H, “Life with 6000 genes” Science, 1996;274:546–67.
- Surya Saha & Susan Bridges & Zenaida V. Magbanua & Daniel G. Peterson, “Computational Approaches and Tools Used in Identification of Dispersed Repetitive DNA Sequences,” Springer Science+Business Media, LLC, 2008
- Alex Van Belkum, Stewart Scherer, Loek Van Alphen, And Henri Verbrugh , “Short-Sequence DNA Repeats in Prokaryotic Genomes,” American Society for Microbiology Microbiology And Molecular Biology Reviews, June 1998, p. 275–293 Vol. 62, No. 2
- Morgante M, Hanafey M, Powell W, “Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes,”
- Nat Genet, 2002; 30:194–200
- Lim S, Notley-McRobb L, Lim M, “A comparison of the nature and abundance of microsatellites in 14 fungal genomes,” Fungal Genet Biol, 2004; 41:1025–36
- Stefan Kurtz, Jomuna V. Choudhuri, Enno Ohlebusch, “REPuter: the manifold applications of repeat analysis on a genomic scale,” Nucleic Acids Research, 2001 Vol 29 No. 22 pg 4633 – 4642
- Chris Abajian, “Sputnik - DNA microsatellite repeat search utility,” http://www.espressosoftware.com/sputnik/
- Schmidt JP ,” All Highest Scoring Paths in Weighted Grid Graphs and Their Application to Finding All Approximate Repeats in Strings,”Siam J. Comput,1998, Vol. 27 Issue 4, 972-992

Thank You

Dr. Mamta C. Padole

mamta.padole@gmail.com

mamta.padole-cse@msubaroda.ac.in