International Journal for Computational Biology (IJCB)
Vol.7, No.1, Apr 2018, pp. 35~48
ISSN: 2278-8115

FrameOUT and FrameOUTDB: A web based application and repository for the identification and analysis of frameshift mutations

Jyoti Thakur, TirathaRaj Singh*
Department of Biotechnology and Bioinformatics, Jaypee University of Information Technology (JUIT), Solan, H.P., India

Article Info

Article history:

Received Jul 10th, 2017
Revised Aug 20th, 2017
Accepted Apr 4th, 2018

Keyword:

FrameOUT (FO)
FrameOUT DB (FODB)
Frameshift
Hidden Markov Model (HMM)

ABSTRACT

Frameshift, one of the three classes of recoding, leads to waste of energy, resources and activity of biosynthetic machinery. In addition, some peptides, probably cycotoxic synthesized after frameshifts, results in diseases and disorders like muscular dystrophies, lysosomal storage disorders, and cancer. Hidden Stop Codons that occur naturally in coding sequences among all organisms, are associated with the early termination of translation for incorrect reading frame selection and help to reduce the metabolic cost related to the frame-shift events. Hidden stop codons and their association with numerous diseases. These codons are associated with the early termination of translation for incorrect reading frame selection and help to reduce the metabolic cost related to the frame-shift events. There are lots of appearances of hidden stops in mitochondrial genomes and we tried to study this putative event in mitochondrial genomes of vertebrates. To reduce this gap, this work presents an algorithmic web based tool to study hidden stops in frame-shifted translation for vertebrate mitochondrial genomes through respective genetic code system. FrameOUT (FO), an algorithmic web based application, predicts mutations in a user input sequence, be it a diseased or a normal sequence by implementation of Hidden Markov Model. FODB is a collection of all available Frameshift events and their association with various diseases.

Corresponding Author:

* Tiratha Raj Singh
Department of Biotechnology and Bioinformatics, Jaypee University of Information Technology

How to Cite:

First Author name *et. al.* FrameOUT and FrameOUTDB: A web based application and repository for the identification and analysis of frameshift mutations. IJCB. 2018; Volume 7 (Issue 1): Page 35-48.

1. INTRODUCTION

Reading frames play an important role in the process of translation of nucleotide sequences into proteins. Selection of a wrong reading frame can alter the protein product. Such events that alter the reading frame are rare during translation; Frame-shift is one such event. Frame-shift is quite common in viruses, bacteria, yeast and other organisms [1, 2]. It is a type of genetic mutation generally caused by indels, i.e. insertion and deletion of nucleotides.

Frame-shifts are defined as protein translations that start not at the first, but either at the second (+1 frame-shift) or the third (−1 frame-shift) nucleotide of the codon [3]. Presumably, most frame-shifts would yield non-functional proteins. Therefore, frame-shifts lead to the waste of energy, resources and activity of the biosynthetic machinery. Some peptides, synthesized after frame-shifts, are probably cytotoxic and serve as possible cause for innumerable diseases and disorders such as muscular dystrophies, lysosomal storage disorders, and cancer. Frame-shift mutations might be beneficial sometime such as a frameshift mutation was

responsible for the creation of Nylonaser [4, 5, and 6]. Coding sequences lack stop codons, but many stop codons appear off-frame. Off-frame stops i.e. stop codons in +1 and −1 shifted reading frames, are termed **Hidden Stop Codons (HSCs)** or hidden stops [7-10].

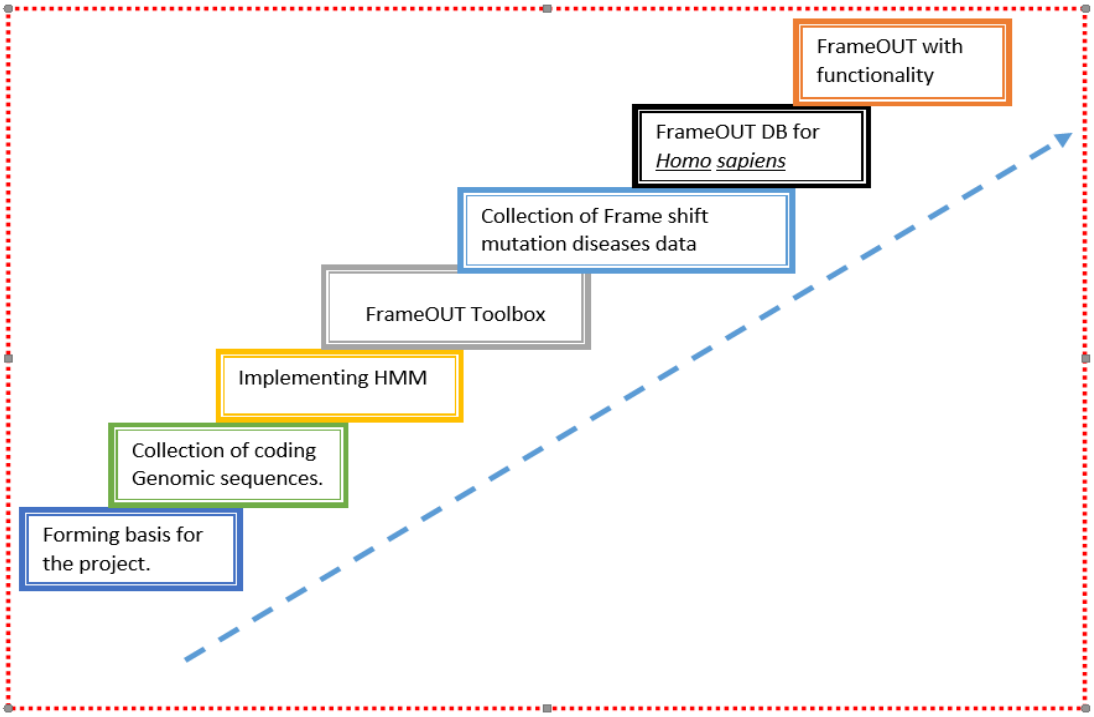
What causes frame-shift errors?

One clear implication of the suppressor analysis is that frame-shifting is strongly stimulated by near-cognate decoding, that is decoding by an isoacceptor that makes a less than optimal wobble interaction with the mRNA. The example of suppression by a structurally normal near cognate tRNA in the*sufB2* strain of *S. typhimurium* clearly shows that near-cognate decoding can stimulate frame errors. Moreover, overproduction of same near-cognate tRNA induces frame-shifting at the same sites suppressed by *sufB2*. Some programmed frame-shifts are also stimulated by near-cognate decoding. The first example comes from the *dnaX* gene of *E.coli*, which encodes alternative forms of a subunit of DNA polymerase III [11]. Frame-shifting results in the expression of a C-terminally truncated form of the protein and occurs on a slippery heptameric sequence A-AAA-AAG, two tRNAs simultaneously slipping −1 from AAA-AAG to AAA-AAA. The unusually high efficiency of this site partly results from the near-cognate recognition of the AAG codon by a tRNA with a modified U in the wobble position which restricts the ability of tRNA to decode AAG. Expressing a tRNA that recognizes AAG in a completely cognate fashion reduced frame-shifting on the site. The weakness of the interaction apparently predisposes the ribosome to frame-shift [2].

The aim of this study was to identify and analyze frame-shift mutations and their disease specific consequences. For this purpose, several mitochondrial vertebrate genomic sequences having 13 protein coding sequences, namely: ND1, ND2, COX1, COX2, ATP6, COX3, ND3, ND4L, ND4, ND5, ND6 and CYTB were collected.
Based on these protein coding sequences, respective transition matrices have been developed. After generating transition matrices, HMM forward algorithm is implemented to compute the joint probability, on the sequence entered by the user, neglecting the stop codons as coding regions lack stop codons [12, 3]. There can be various probable states such as 0, 1 and 2 that are based upon the position of characters in nucleotide sequences for normal translation (0), +1 frameshift -1 and 1frameshift (2), and symbols are our very own nucleotides i.e. A, T, G and C, with equal probabilities.
Therefore, FrameOut (FO) is a web based tool that predicts the mutational events occurring in genomic sequences through frame-shift events. Data is being framed through HMM. The calculation of probability of mutation in the user input sequence is done by implementing Hidden Markov Model (which is equivalent to stochastic regular grammars), particularly HMM Forward Algorithm, in which we calculate the probability based on certain training set.
FrameOut DB (FODB) is a collection of all available Frameshift events and their association with various diseases specifically human diseases such as Corhn's disease, Rett-Syndrome, and Sandhoff disease, etc [13, 14 and 15].

2. METHODOLOGY

The entire research work was carried out in phases. The figure below narrates the entire process.



The FrameOut (FO) tool, that is a web based tool to predict the mutational events occurring in genomic sequences through frame-shift events, and FrameOUT DB have been designed with the help of languages such as HTML, CSS, JavaScript and PHP, MySQL and WAMP server.

Steps to create FrameOUT Tool:

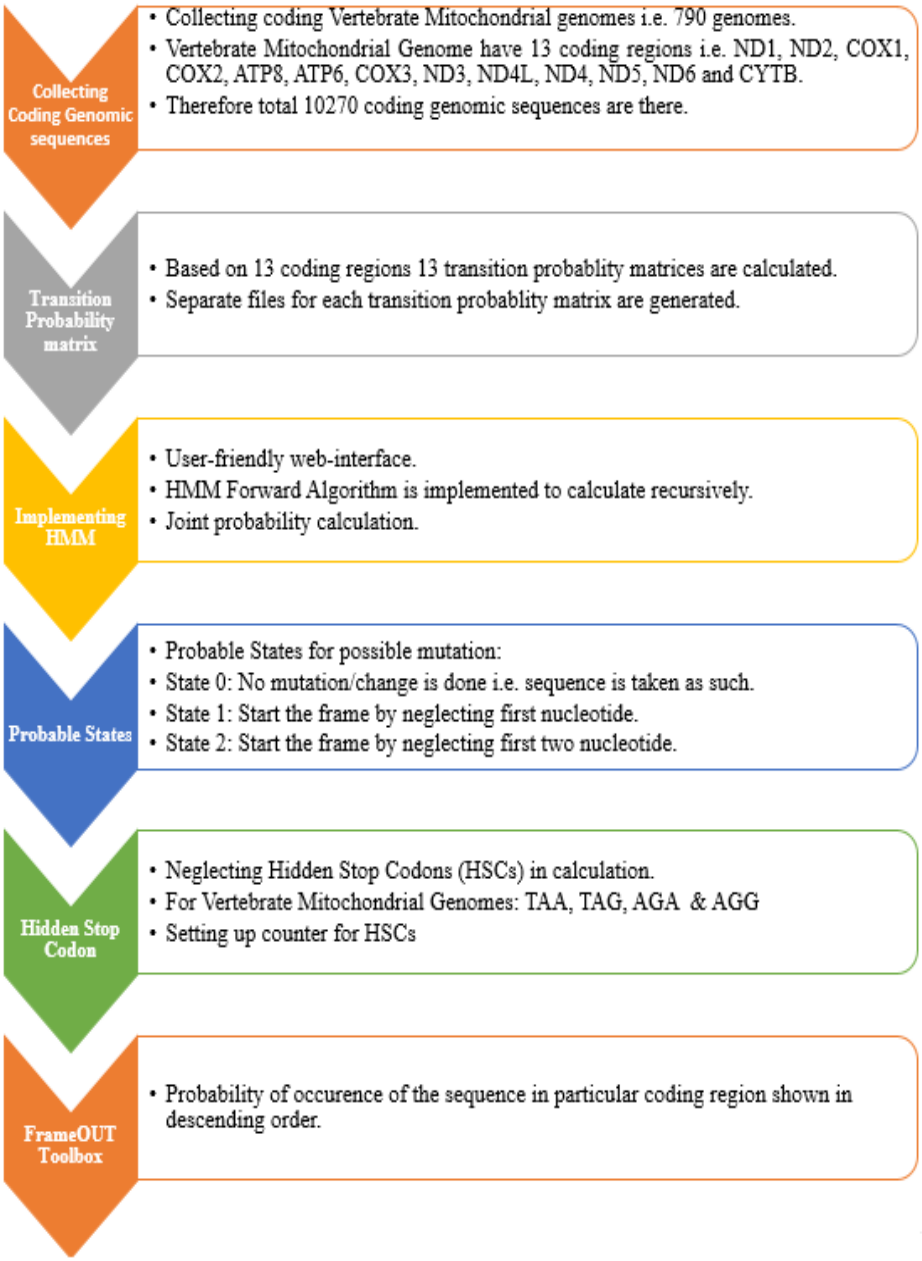
1. Collection of Coding Genomic sequences
2. Generation of Transition Probability matrix
3. Algorithmic Implementation of HMM
4. Identification of Probable States
5. Identification and analysis of Hidden Stop Codon

Steps to create FrmaOUT DB:

- Collecting Data
- Enlisting Related Attributes
- Linking with Other DB
- Various Options and their implementations
- FODB development

Both the processes are explained through the following diagrams.

Methodology for FrameOut Tool:



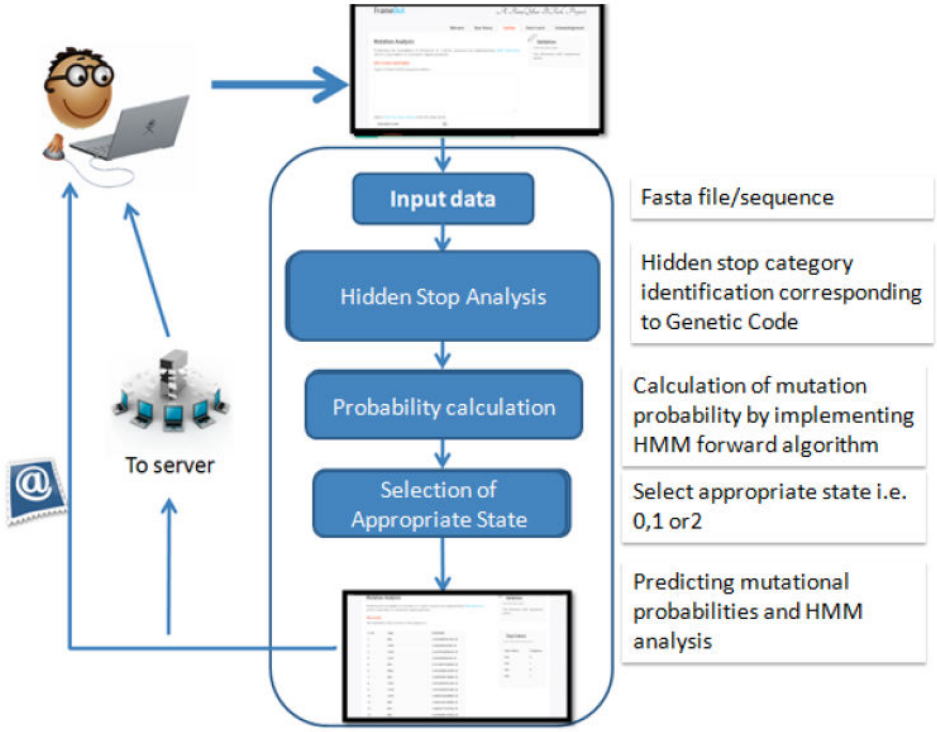
Methodology for FrameOUT DB:

Data Collection

While collecting coding Genomic sequences, particularly, Vertebrate mitochondrial genomes, we collected genomic sequence data for 790 genomes. The collected genomes had 13 protein coding regions, namely: ND1, ND2, COX1, COX2, ATP6, ATP8, COX3, ND3, ND4L, ND4, ND5, ND6 and CYTB. After collecting the genomic sequence data, I segregated these 790 genomes into respective coding sequence files. Therefore, separate files listing respective coding regions have corresponding sequences and thereby there are 13 files, each having different 790 sequences. So, in total, there were 10270 coding genomic sequences.

In this section, the basic flowchart of how FrameOUT works is described through the diagram below.

Basic Flowchart



3. RESULTS

Transition Probability Calculation

For each separate file transition probability matrix was calculated, therefore 13 different transition probability matrices were generated based on the following formula:

$$Transition\ probability = p(x_t|x_{t-1}) = p(y|x) = \frac{p(xy)}{p(x)} \approx \frac{freq(xy)}{freq(x)}$$

$$= \frac{Frequency\ of\ particular\ codon}{Frequency\ of\ all}$$

For example: -= $$\frac{Frequency\ (AAA)}{Frequency\ (AAA+AAT+AAG+AAC)}$$

Transition Probability Matrix for ND1

	AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT	Total
A	0.057822	0.066803	0.028136	0.07472	0.079752	0.07522	0.031846	0.118937	0.10145771	0.101073	0.055142	0.071593	0.077195	0.076719	0.050524	0.0898291	1
C	0.091784	0.08985	0.066661	0.077978	0.05887	0.084142	0.030344	0.077563	0.06441337	0.149361	0.067974	0.044816	0.081532	0.087119	0.036541	0.08343339	1
G	0.032106	0.031763	0.035119	0.048846	0.041684	0.032959	0.018576	0.036603	0.05823931	0.018147	0.033395	0.022773	0.058943	0.080298	0.025083	0.03742348	1
T	0.083519	0.081855	0.031905	0.101133	0.078972	0.131991	0.020379	0.082161	0.05379123	0.078556	0.03234	0.046927	0.105164	0.075381	0.023343	0.06147234	1

Transition Probability Matrix for ND2

	AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT	Total
A	0.057835	0.066839	0.028115	0.074714	0.079789	0.075221	0.031803	0.118962	0.10143341	0.101147	0.055156	0.071552	0.077217	0.076742	0.050526	0.08981282	1
C	0.091809	0.089856	0.066667	0.077977	0.058883	0.08415	0.030343	0.077582	0.06438979	0.149408	0.067981	0.044798	0.08153	0.087157	0.036543	0.08344647	1
G	0.032099	0.031722	0.035112	0.048826	0.041681	0.032958	0.018568	0.036559	0.05822408	0.018142	0.033374	0.022818	0.058936	0.030294	0.025064	0.03741005	1
T	0.083521	0.081853	0.031879	0.101176	0.078981	0.132009	0.020371	0.082141	0.05381436	0.078566	0.032328	0.046869	0.105193	0.07536	0.023324	0.06141574	1

Transition Probability Matrix for COX1

	AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT	Total
A	0.057797	0.066888	0.028116	0.074723	0.07976	0.075199	0.03181	0.118964	0.1013389	0.101033	0.055241	0.071497	0.077226	0.076751	0.050548	0.08977329	1
C	0.091797	0.089855	0.066672	0.077964	0.058914	0.084142	0.030356	0.077508	0.06443286	0.149299	0.068009	0.044807	0.081575	0.087124	0.036526	0.08342598	1
G	0.032088	0.03171	0.035094	0.048819	0.041639	0.032983	0.018631	0.036614	0.05828514	0.018197	0.033379	0.022818	0.058904	0.030316	0.02513	0.03750104	1
T	0.08352	0.081941	0.031858	0.101158	0.079023	0.131956	0.020362	0.082139	0.05384183	0.078561	0.032374	0.046886	0.105088	0.075267	0.023381	0.06146507	1

Transition Probability Matrix for COX2

	AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT	Total
A	0.057796	0.066933	0.028162	0.074707	0.079748	0.075265	0.031835	0.118859	0.10150269	0.101	0.055289	0.071577	0.077209	0.076631	0.050566	0.08980006	1
C	0.091795	0.089808	0.066629	0.077919	0.059047	0.084038	0.030361	0.077589	0.064512	0.149145	0.067882	0.044795	0.081476	0.087125	0.036549	0.08345051	1
G	0.032096	0.031748	0.035097	0.048825	0.041655	0.032948	0.018612	0.036698	0.05822535	0.01815	0.033355	0.022841	0.058915	0.030376	0.025163	0.03744078	1
T	0.08354	0.081936	0.031831	0.101177	0.07894	0.131866	0.020419	0.082118	0.05382076	0.078583	0.032419	0.046903	0.105159	0.075347	0.023365	0.06142796	1

Transition Probability Matrix for ATP8

	AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT	Total
A	0.057806	0.066889	0.028186	0.07473	0.079689	0.075225	0.031857	0.118965	0.10154878	0.100986	0.055214	0.071589	0.077246	0.076574	0.050561	0.08980691	1
C	0.091766	0.089783	0.066658	0.077936	0.058982	0.084125	0.030377	0.07757	0.06452599	0.149157	0.067929	0.044747	0.081553	0.087222	0.036539	0.08338197	1
G	0.032109	0.031741	0.035025	0.048861	0.041639	0.032999	0.018635	0.036649	0.05814343	0.018191	0.033333	0.022847	0.058952	0.030315	0.025155	0.03746083	1
T	0.083542	0.081986	0.031854	0.101129	0.078923	0.131901	0.020359	0.082104	0.05385122	0.078672	0.032376	0.046888	0.105153	0.075299	0.02339	0.06139004	1

Transition Probability Matrix for ATP6

	AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT	Total
A	0.057776	0.066921	0.028171	0.074759	0.079814	0.075291	0.031858	0.118893	0.10153181	0.101043	0.055291	0.071606	0.077226	0.076505	0.050531	0.08976434	1
C	0.091801	0.089773	0.066631	0.077971	0.058962	0.084045	0.030353	0.077515	0.0645295	0.149125	0.067846	0.04475	0.081459	0.087209	0.036536	0.08351483	1
G	0.032085	0.031717	0.035027	0.048852	0.041661	0.033033	0.018609	0.036578	0.05818315	0.018082	0.03338	0.022849	0.058924	0.030335	0.02516	0.03751668	1
T	0.083614	0.08189	0.031874	0.101137	0.078925	0.131804	0.020395	0.082263	0.05387986	0.078594	0.032363	0.046941	0.105128	0.075367	0.023379	0.06144533	1

Transition Probability Matrix for COX3

	AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT	Total
A	0.05773	0.06694	0.028165	0.074696	0.07987	0.075299	0.031835	0.118989	0.10148561	0.101091	0.05516	0.071546	0.077261	0.076559	0.050573	0.08978033	1
C	0.091824	0.0898	0.066601	0.077957	0.058982	0.084032	0.03034	0.077481	0.06447855	0.149124	0.068072	0.044786	0.08154	0.087223	0.036547	0.08342494	1
G	0.032104	0.031711	0.035052	0.048898	0.041673	0.033025	0.018579	0.036585	0.058161	0.018045	0.033384	0.022832	0.058926	0.0303	0.025191	0.03751889	1
T	0.083635	0.081887	0.031883	0.101117	0.078948	0.131831	0.020375	0.082155	0.05386704	0.078713	0.032318	0.046937	0.105097	0.075278	0.02338	0.06139946	1

Transition Probability Matrix for ND3

	AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT	Total
A	0.057717	0.066975	0.02819	0.074782	0.079843	0.075269	0.031822	0.118941	0.10148865	0.101024	0.055209	0.071558	0.077192	0.076573	0.050582	0.08975983	1
C	0.091812	0.089855	0.066577	0.078007	0.058998	0.084013	0.030353	0.077476	0.06451454	0.148944	0.067982	0.044837	0.081582	0.087191	0.036525	0.08340946	1
G	0.032094	0.031735	0.035024	0.048868	0.04164	0.03297	0.018605	0.03666	0.05810358	0.018146	0.033408	0.022868	0.058946	0.030288	0.025135	0.0375178	1
T	0.083577	0.081871	0.031872	0.101045	0.078982	0.131842	0.020375	0.082211	0.05395474	0.07864	0.032371	0.046951	0.105064	0.075325	0.023427	0.06142134	1

Transition Probability Matrix for ND4L

	AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT	Total
A	0.057718	0.067002	0.028173	0.074712	0.079844	0.07534	0.031842	0.118946	0.10150658	0.100933	0.055291	0.071563	0.07721	0.076576	0.05064	0.08969216	1
C	0.091736	0.089797	0.066579	0.078033	0.059054	0.084032	0.030339	0.07745	0.06467239	0.149057	0.067925	0.044753	0.081556	0.087126	0.036506	0.08349802	1
G	0.032089	0.03174	0.035021	0.048916	0.041627	0.032949	0.018632	0.036723	0.05816758	0.018061	0.033404	0.022806	0.058923	0.030376	0.02517	0.03746964	1
T	0.083691	0.081881	0.031848	0.101064	0.078996	0.13176	0.020385	0.082081	0.05392661	0.078532	0.032336	0.047066	0.105025	0.075337	0.023449	0.06144491	1

Transition Probability Matrix for ND4

	AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT	Total
A	0.057677	0.06702	0.028157	0.074775	0.079859	0.075287	0.031823	0.119005	0.10146675	0.100893	0.055329	0.071568	0.077296	0.076559	0.050568	0.08971941	1
C	0.091785	0.089739	0.066543	0.077991	0.059055	0.08406	0.03033	0.077472	0.06451135	0.148909	0.068	0.044882	0.081604	0.08714	0.036528	0.08341684	1
G	0.032072	0.031737	0.035003	0.048821	0.041613	0.032934	0.01861	0.036716	0.05822527	0.018166	0.033377	0.022762	0.05887	0.03028	0.0252	0.03755993	1
T	0.083701	0.081955	0.031821	0.101162	0.078965	0.13171	0.02038	0.082181	0.05402467	0.078744	0.03227	0.046869	0.105007	0.075375	0.023427	0.06144875	1

Transition Probability Matrix for ND5

	AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT	Total
A	0.05769	0.067086	0.028199	0.074785	0.079967	0.075409	0.031733	0.118934	0.10143253	0.100959	0.055317	0.071469	0.077301	0.076569	0.05059	0.08977955	1
C	0.091792	0.089705	0.06653	0.07809	0.05911	0.083975	0.030354	0.077483	0.06450849	0.148911	0.067974	0.044595	0.081467	0.087267	0.036555	0.08341022	1
G	0.032059	0.031657	0.035001	0.04887	0.041611	0.032946	0.018623	0.036686	0.05817019	0.018097	0.033419	0.022905	0.058947	0.03019	0.025238	0.03754613	1
T	0.083772	0.081872	0.031824	0.101118	0.078998	0.131729	0.020283	0.082159	0.05403351	0.078794	0.032412	0.047004	0.105037	0.075274	0.023398	0.0614302	1

Transition Probability Matrix for ND6

	AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT	Total
A	0.057681	0.06702	0.028165	0.074742	0.079894	0.075432	0.031809	0.118938	0.10132167	0.100749	0.055275	0.071394	0.077311	0.076531	0.050628	0.08968101	1
C	0.0517	0.089674	0.066582	0.078074	0.059165	0.083928	0.030331	0.077456	0.06464254	0.148977	0.068058	0.044934	0.081377	0.087211	0.036603	0.08343432	1
G	0.032048	0.031718	0.035059	0.048853	0.041623	0.032859	0.018616	0.036794	0.05809833	0.018132	0.033511	0.022979	0.058966	0.030264	0.025309	0.03759993	1
T	0.083645	0.081898	0.031817	0.101127	0.078906	0.131768	0.020305	0.082177	0.05398229	0.078728	0.032474	0.046925	0.104582	0.075298	0.023453	0.06135216	1

Transition Probability Matrix for CYTB

	AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT	Total
A	0.057718	0.066974	0.028161	0.07478	0.079836	0.075431	0.031805	0.118879	0.10134302	0.1008	0.05523	0.071532	0.077296	0.076417	0.050492	0.08963517	1
C	0.091813	0.089728	0.066557	0.07758	0.058976	0.08401	0.03038	0.077591	0.06444705	0.148857	0.068078	0.04481	0.081438	0.087235	0.036513	0.0834949	1
G	0.032044	0.031656	0.035068	0.048549	0.041673	0.032892	0.01855	0.036746	0.05805267	0.018137	0.033551	0.022795	0.058895	0.030343	0.025391	0.03759749	1
T	0.083761	0.081868	0.031838	0.101103	0.07897	0.131748	0.020324	0.082189	0.05390817	0.078746	0.032584	0.047129	0.10492	0.075411	0.023506	0.06139778	1

Transition Probability Matrix for ND4

	AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT	Total
A	0.057677	0.06702	0.028197	0.074775	0.079859	0.075287	0.031823	0.119005	0.10146675	0.100893	0.055329	0.071568	0.077296	0.076559	0.050568	0.08971941	1
C	0.091785	0.089739	0.066543	0.077991	0.059055	0.08406	0.03033	0.077472	0.06451135	0.148909	0.068	0.044882	0.081604	0.08714	0.036528	0.08341684	1
G	0.032072	0.031737	0.035063	0.048821	0.041613	0.032934	0.01861	0.036716	0.05822527	0.018166	0.033377	0.022762	0.05887	0.03028	0.0252	0.03755993	1
T	0.083701	0.081955	0.031821	0.101162	0.078965	0.13171	0.02038	0.082181	0.05402467	0.078744	0.03227	0.046869	0.105017	0.075375	0.023427	0.06144875	1

Transition Probability Matrix for ND5

	AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT	Total
A	0.05769	0.067086	0.028199	0.074785	0.079967	0.075409	0.031733	0.118934	0.10143253	0.100959	0.055317	0.071469	0.077301	0.076569	0.05059	0.08977995	1
C	0.091792	0.089705	0.06653	0.07809	0.05911	0.083975	0.030354	0.077483	0.06450849	0.148911	0.067974	0.044595	0.081467	0.087267	0.036555	0.08341022	1
G	0.032059	0.031657	0.035001	0.04887	0.041611	0.032946	0.018623	0.036686	0.05817019	0.018097	0.033419	0.022905	0.058947	0.03019	0.025238	0.03754613	1
T	0.083772	0.081872	0.031824	0.101118	0.078998	0.131729	0.020283	0.082159	0.05403351	0.078794	0.032412	0.047004	0.105037	0.075274	0.023358	0.0614302	1

Transition Probability Matrix for ND6

	AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT	Total
A	0.057681	0.06702	0.028165	0.074742	0.079894	0.075432	0.031809	0.118938	0.10132167	0.100749	0.055275	0.071394	0.077311	0.076531	0.050628	0.08968101	1
C	0.0917	0.089674	0.066582	0.078074	0.059165	0.083928	0.030331	0.077456	0.06464254	0.148977	0.068058	0.044694	0.081377	0.087211	0.036603	0.08343482	1
G	0.032048	0.031718	0.035056	0.048853	0.041623	0.032859	0.018616	0.036794	0.05809833	0.018192	0.033511	0.022979	0.058966	0.030264	0.025309	0.03759993	1
T	0.083645	0.081898	0.031817	0.101127	0.078906	0.131768	0.020305	0.082177	0.05398229	0.078728	0.032474	0.046925	0.104982	0.075298	0.023453	0.06135216	1

Transition Probability Matrix for CYTB

	AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT	Total
A	0.057718	0.066974	0.028161	0.07478	0.079836	0.075431	0.031805	0.118879	0.10134302	0.1008	0.05523	0.071532	0.077296	0.076437	0.050492	0.08963517	1
C	0.091813	0.089728	0.066557	0.07798	0.058976	0.08401	0.03038	0.077591	0.06444705	0.148857	0.068078	0.04481	0.081438	0.087235	0.036513	0.0834949	1
G	0.032044	0.031656	0.035068	0.048545	0.041673	0.032892	0.01855	0.036746	0.05805267	0.018137	0.033551	0.022795	0.058895	0.030343	0.025391	0.03756749	1
T	0.083761	0.081868	0.031838	0.101103	0.07897	0.131748	0.020324	0.082189	0.05390817	0.078746	0.032584	0.047129	0.10492	0.075411	0.023506	0.06139778	1

Implementing HMM: Forward Algorithm

Implementing HMM forward algorithm, which basically computes the joint probability*p (xt, y1: t)*, neglecting the stop codons. The forward algorithm takes advantage of the conditional independence rules of HMM to perform the calculation recursively based on following formula [16, 17]:

$$\alpha_t(x_t) = p(y_t|x_t)\sum_{x_{t-1}} p(x_t, x_{t-1})\alpha_{t-1}(x_{t-1})$$

Thus, since

- SYMBOLS = A, T, G, C
- STATE = 0, 1, 2

$p(y_t|x_t)$= *emission distributions probability*= 0.25 i.e. equal probability for each SYMBOL

$p(x_t|x_{t-1})$ = *transition probabilities* = value from transition probability matrix calculated previously

$\alpha_{t-1}(x_{t-1})$ = *previously calculated probability*

The various probable STATES are 0, 1 and 2:

- State 0: no mutation/change is done i.e. sequence is taken as such,
- State 1: start the frame by neglecting first nucleotide and
- State 2: start the frame by neglecting initial two nucleotides

FrameOUT Tool Box

FrameOUT DB

FODB is a collection of all available frame-shift events and their association with various diseases specifically human diseases such as Corhn's disease, Rett-Syndrome, and Sandhoff disease, etc. This database has various options which may aid to give better results. Also, certain attributes are linked with other databases/resources such as gene list is linked with GenBank and PMID with PubMed.

FrameOUT DB

Availability of the tool and System Requirements:

- **FrameOUT home page:** http://www.bioinfoindia.org/frameout
- **FrameOUT Tool:** http://bioinfoindia.org/frameout/toolbox.php
- **FrameOUT DB:** http://bioinfoindia.org/frameout/frameoutdb.php
- **Programming Languages:** PHP 5.3.13 / HTML 4.0
- **Web Server:** Apache 2.2 through WampServer
- **Database Server:** MySQL 3.5.1
- **Other requirements:** Web enabled services from standard web browsers

4. CONCLUSION

A new algorithmic tool, FrameOUT, has been developed, which allows user-friendly exploration, analysis, and visualization of mutational probabilities and hidden stop codons with the mitochondrial vertebrate genome analysis. This web based tool is perfect to serve as a useful complement for analyzing hidden stop codons in all available genetic code systems, particularly for vertebrate mitochondrial genetic code.
FramOUT DB is a collection of diseases caused to frame-shift mutational events such as Corhn's disease, Rett-Syndrome, and Sandhoff disease, etc. User can make specific searches by using various search options enlisted in the database. It is anticipated that the developed tool and resource on frameshift mutations will provide insightful information to the biotechnologists and the biomedical scientists.It will also provide an opportunity for the analysis of existing data as well as support and exploratory analysis for newly generated data.

AVAILABILITY

- **FrameOUT home page:** http://www.bioinfoindia.org/frameout
- **FrameOUT Tool:** http://bioinfoindia.org/frameout/toolbox.php
- **FrameOUT DB:** http://bioinfoindia.org/frameout/frameoutdb.php

REFERENCES

1. Belshaw, R., Pybus, O.G., Rambaut, A., 2007. The evolution of genome compression and genomic novelty in RNA viruses. Genome Res. 17, 1496–1504.
2. Baranov PV, Gesteland RF, Atkins JF (2002) Release factor 2 frame-shifting sites in different bacteria. EMBO Rep 3: 373-377.
3. Louise J Johnson**,** James A Cotton**,** Conrad P Lichtenstein**,** Greg S Elgar**,** Richard A Nichols**,** p David Polly and Steven C Le Comber (2011) Stops making sense: Translational trades-offs and stop codon reassignment. *BMC Evolution Biology* 11:227.
4. Gupta A, Singh TR (2013) SHIFT: Server for hidden stops analysis in frame-shifted translation. BMC Research Notes 6:68.
5. Ohno S: Birth of a unique enzyme from an alternative reading frame of the pre-existed, internally repetitious coding sequence. Proc Natl Acad Sci USA 1984, 81:2421–2425.
6. J.F., Giddings, M.C., 2001. RECODE: a database of frame-shifting, bypassing and codon redefinition utilized for gene-expression. Nucleic Acids Res. 296, 264–267.
7. Singh TR, Pardasani KR (2009) Ambush hypothesis revisited: Evidences for phylogenetic trends. Computational Biology Chem 33: 239-244.
8. Singh TR (2013) Mitochondrial Genomes and Frameshift Mutations: Hidden Stop Codons, their Functional Consequences and Disease Associations Journal of Clinical & Medical Genomics 1: 108. doi: 10.4172/ijgm.1000108.
9. Doug A. Brooks, Viv J. Muller, John J. Hopwood (2006) Stop-codon read-through for patients affected by a lysosomal storage disorder. *Trends in Molecular Medicine* Vol.12 No.8.doi.org/10.1016/j.molmed.2006.06.001.
10. Louise J Johnson**,** James A Cotton**,** Conrad P Lichtenstein**,** Greg S Elgar**,** Richard A Nichols**,** p David Polly and Steven C Le Comber (2011) Stops making sense: Translational trades-offs and stop codon reassignment. *BMC Evolution Biology* 11:227.
11. Martina MA, Correa EME, Argaraña CE, Barra JL: *Escherichia coli* Frameshift Mutation Rate Depends on the Chromosomal Context but Not on the GATC Content Near the Mutation Site. PLoS ONE 2012, 7: e33701.
12. Russell RD, Beckenbach AT (2008) Recoding of translation in turtle mitochondrial genomes: programmed frame-shift mutations and evidence of a modified genetic code. J Mol Evol 67: 682-695.
13. Tse H, Cai JJ, Tsoi H-W, Lam EPT, Yuen K-Y: Natural selection retains overrepresented out-offrame stop codons against frame-shift peptides in prokaryotes. BMC Genomics 2010, 11:491.
14. Littink KW, van Genderen MM, van Schooneveld MJ, Visser L, Riemslag FC, Keunen JE, Bakker B, Zonneveld MN, den Hollander AI, Cremers FP, van den Born LI: A Homozygous Frameshift Mutation in LRAT Causes Retinitis Punctata Albescens. Ophthalmology 2012, 119:1899–1906.
15. Sagong B, Seok JH, Kwon TJ, Kim UK, Lee SH, Lee KY: A novel insertion induced frame-shift mutation of the SLC26A4 gene in a Korean family with Pendred syndrome. Gene 2012, 508:135–139.
16. Richard Durbin, Sean R. Eddy, Anders Krogh & Graeme Mitchison *Biological sequence analysis*: Probabilistic models of proteins and nucleic acids.
17. The GRAMMAR package, The *MARKOV package: Markovian Model*, https://www.lri.fr/~genrgens/manual/GRGs-manual-html/node4.html(Accessed: 2013-2014).