BIOINFORMATION
Discovery at the interface of physical and biological sciences

Open access

BIOMEDICAL INFORMATICS

www.bioinformation.net
Volume 12(4)

Prediction Model

Prediction of miRNA binding sites in mRNA

Anatoly Ivashchenko, Anna Pyrkova, Raigul Niyazova*, Aigul Alybayeva, Kirill Baskakov

Computer Science Laboratory, Al-Farabi Kazakh National University, Almaty - 050038, Kazakhstan, Raigul Niyazova – Email: raiguln@mail.ru, *Corresponding author

Recieved May 26, 2016; Revised July 4, 2016; Accepted July 4, 2016; Published July 26, 2016

Abstract:
In the present article, a solution for the problem of gene scanning for the purpose of prediction of binding sites of miRNA with matrix RNA (mRNA) is proposed. A mathematical model for the optimal process of scanning of genes and miRNA sequences was developed. An algorithm for scanning of genes using miRNA with one gap in the miRNA sequence and maximum (as a percentage ratio) free energy was developed and analysed with regard to the coincidence of miRNA and particular gene sites on the basis of complementarity. The constructed algorithm for scanning of genes using miRNA was parallelised and fragmented on the computational cluster with the use of MPJ tools.

Keywords: parallelized algorithm, cluster computing platform, Java MPI, miRNA, mRNA.

Background:
Searching for miRNA binding sites is difficult **[1]**. It is necessary to take into account the many characteristics (e.g., binding energy, degree of complementarity of nucleotides in the interaction of miRNA with mRNA, presence of non-interacting nucleotides, and multiplicity of binding sites located using one or more nucleotides). Recently, the number of identified miRNAs has increased to over 3,000 **[2]**. The up-to-date database mirBase.org contains approximately about 2,700 miRNA sequences. The volume of calculations for finding miRNA binding sites has increased significantly, and therefore requires the development of computer technology to increase the speed of calculations. In this study, we developed a program for prediction of miRNA binding sites that allowed us to increase by dozens of fold the speed of detection of miRNA binding sites. This program uses new additional characteristics for the interaction of nucleotides in the miRNA binding sites to the mRNA. Scanning of a genome enables hundreds of possible targets to be revealed for therapy of various diseases. Such scanning is important because knowledge of the interacting genes will allow the roles of the protein to be defined and will enable identification of intracellular processes that are disrupted in the disease.

Methodology:
Mathematical model of scanning genes
Scanning of genes is a process of consecutive comparison of sites of a gene with miRNA with possibility of adding one gap in the miRNA in positions with the 3rd on n-2-th, where n indicates the nucleotide number (length) of miRNA. Thus, there is an assessment of all possible comparisons at one site of mRNA with miRNA that is defined according to the free energy value of the compared sequences. The option that is closer (in a percentage ratio) with regard to the free energy for coincidence of miRNA and a gene site is considered to be the best on the basis of complementarity. The mathematical model of the problem of scanning of genes can be formulated as follows:

Let $\{u_l\}, l = \overline{1, N}$ be a set of nucleotide or amino-acid sequences of *miRNA*, *N* be the amount of sequences of *miRNA*, $\{v_g\}$, $g = \overline{1, M}$ be a set of sequences of genes of *miRNA*, and M be the amount of sequences of *miRNA*; then $\{\overline{u}, \overline{v}$ $Number, Position, where, energy, score, length\}$ – scanning genes, where *Number* is the order number, *Position* is a position of $\{u_l\}$, in $\{v_g\}$, where is an element from a set $\{5'UTR, CDS, 3'UTR\}$ defining site arrangement area u_l in v_g, $Energy$ is the value of free energy on the basis of a complementarity, *Score* is the value of ΔG/ΔGm, and $Length$ is the length of u_l.

BIOINFORMATION
Discovery at the interface of physical and biological sciences

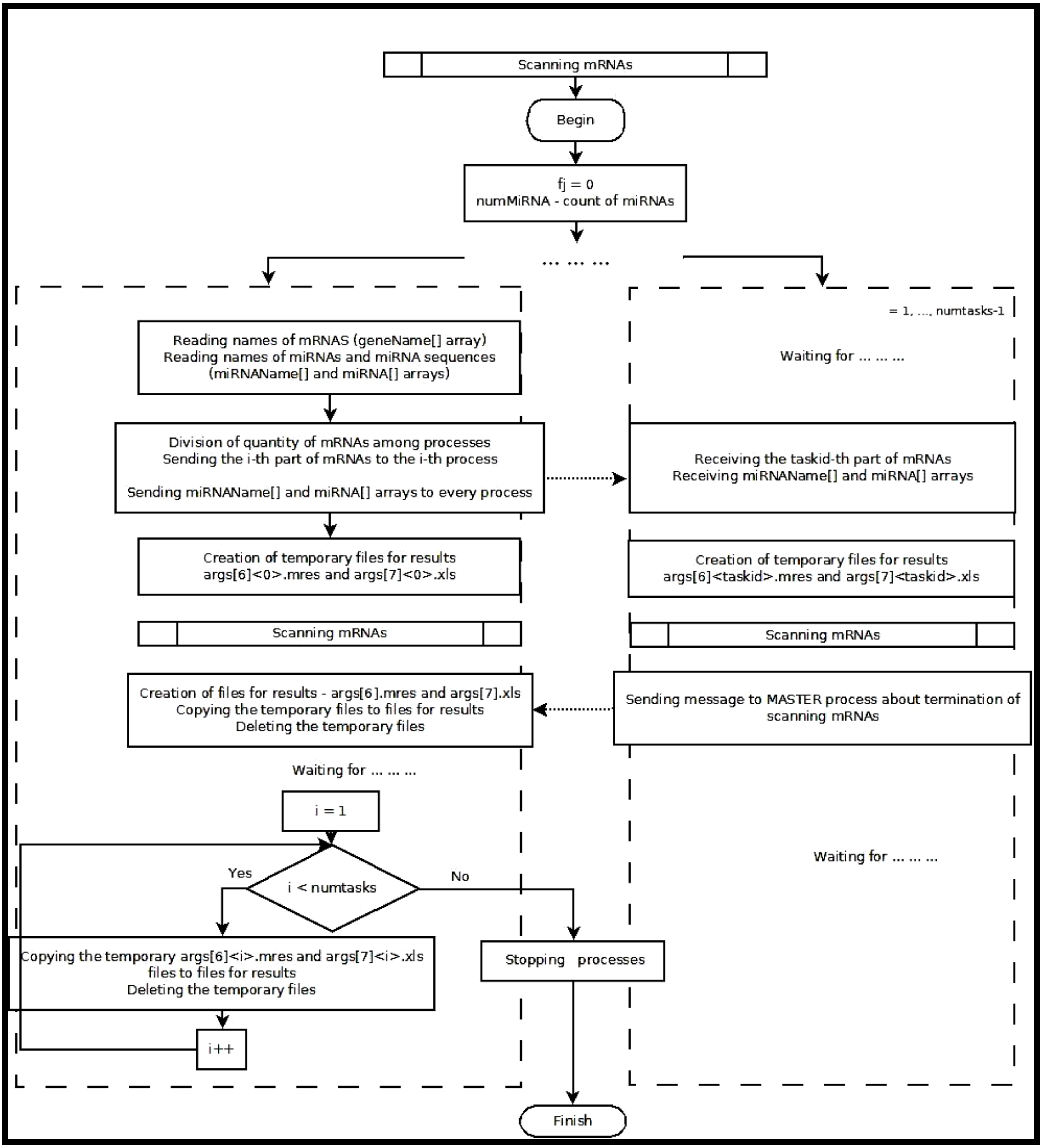

Figure 1: Flowchart of the scanning genes algorithm.

Table 1: Comparative analysis of the linear, parallelized, and fragmented algorithms for scanning genes (time is specified in seconds)

	Nodes	Duration of processing (Average length = 21)		
Number of miRNA		100	1000	10000
Linear algorithm		3	54	532
The parallelized algorithm	15	1	3	8
The fragmented algorithm	15	1	7	53

	A	B	C	D	F	G	H	I	J	K	L	M	N
1	Target gene	miRNA	Position	mRNA domain	Energy	Score	Length	Binding site	mRNA fragment	Oligopeptide	ORF1	ORF2	ORF3
2	TBC1D10B	miR-762(RP	228	CDS	-136	100	22	gcuccggcc	ucggccgagac	SAETSAPAPA	APAPAPA_	LRPRPQP	SGPGPSP
3	BSN	miR-762(RP	184	CDS	-132	96,87	22	gcccgggcc	ggcggcgccgg	GGAGPGPGPC	ARAPAPA_	PGPRPRP	PGPGPGP
4	CDKN1C	miR-762(RP	876	CDS	-132	96,87	22	gccccggcc	gcggccccggc	AAPAPAPAPA	APAPAPA_	PRPRPRP	PGPGPGP
5	CDKN1C	miR-762(RP	882	CDS	-132	96,87	22	gccccggcc	ccggccccagc	PAPAPAPAPA	APAPAPA_	PRPRPRP	PGPGPGP
6	STK39	miR-762(RP	283	CDS	-132	96,87	22	gccccggcc	cccggcagcuc	PGSSRGPGPC	APAPAPA_	PRPRPRP	PGPGPGP
7	TAF4	miR-762(RP	524	CDS	-132	96,87	22	gcccggccc	gccggccccgg	AGPGPGPGPC	ARPRPRP	PGPGPGP	PAPAPA_
8	CDKN1C	miR-762(RP	858	CDS	-129	95,31	22	gucgcggcc	ccggccccggc	PAPAPVAAPA	VAAPAPA_	SRPRPQP	RGPGPSP
9	CDKN1C	miR-762(RP	888	CDS	-129	95,31	22	gccccggcc	ccagccccggc	PAPAPAPAPA	APAPAPA_	PRPRPPP	PGPGPRP
10	RELA	miR-762(RP	1343	CDS	-129	95,31	22	gcucuggcc	cuccagccaug	LQPWYQLWPF	ALAQAPA_	LWPRPQP	SGPGPSP
11	SEMA3G	miR-762(RP	61	CDS	-129	95,31	22	gcucuggcc	cugcuccaugg	LLHGGSSGPS	ALAPAPA_	LWPQPRP	SGPSPGP
12	STARD8	miR-762(RP	1657	CDS	-129	95,31	22	gcugaggcc	aguggcacagg	SGTGRGOGPC	AEAPAPA_	LRPRPQP	OGPGPSP
13	TBC1D10B	miR-762(RP	240	CDS	-129	95,31	22	gccccagcc	ucugcuccggc	SAPAPAPAPA	APAPAPA_	PQPRPQP	PSPGPSP
14	UBE2O	miR-762(RP	107	CDS	-129	95,31	22	gcccaggcc	ccgcagcuccc	PQLPLQPRPF	AQAPAPA_	PRPRLQP	PGPGSSP
15	PRRC2C	miR-762(RP	5990	CDS	-127	93,75	22	gcuucagcc	cacuugccccu	HLPLPQLQPC	ASAPAPA_	LQPQPQP	FSPSPSP
…	…	…	…	…	…	…	…	…	…				
94	SH3BP1	miR-762(RP	1644	CDS	-123	90,62	22	gcuccggcu	ccagccaccac	PATTPAPAPA	APAPAPA_	LRLRLQL	SGSGSSS
95	SHANK3	miR-762(RP	3715	CDS	-123	90,62	22	gcccagccc	cuguggcagag	LWQSPCPAPC	AQPPGPA_	PSPRAQP	PAPGPSP
96	SMARCA4	miR-762(RP	1051	CDS	-123	90,62	22	gcccuggcc	uccgcaacagg	SATGPGPGPC	ALALALA_	PWPWPWP	PGPGPGP
97	STK39	miR-762(RP	278	CDS	-123	90,62	22	ccgcggccc	gccggccccgg	AGPGSSRGPC	PRPRPRP	RGPGPGP	AAPAPA_
98	SULT2B1	miR-762(RP	1170	CDS	-123	90,62	22	acuccagcc	gugagcccaga	VSPDPTPAPA	TPAPAPA_	LQPQPQP	SSPSPSP
99	TAF4	miR-762(RP	505	CDS	-123	90,62	22	gcgccggcc	gccgcgcuggc	AALAARAGPC	APAPAPG_	RRPRPRA	AGPGPGP
100	TAF4	miR-762(RP	535	CDS	-123	90,62	22	gccccggcc	cccgggcccgg	PGPGPGPGPC	APAPALA_	PRPRPWP	PGPGPGP
101	TBC1D10B	miR-762(RP	229	CDS	-123	90,62	22	cuccggccc	ucggccgagac	SAETSAPAPA	LRPRPQP	SGPGPSP	PAPAPA_
102	TBC1D10B	miR-762(RP	241	CDS	-123	90,62	22	ccccagccc	ucugcuccggc	SAPAPAPAPA	PQPRPQP	PSPGPSP	PAPAPA_
103	TNFRSF21	miR-762(RP	1988	CDS	-123	90,62	22	gcccgagcc	gagccccagcc	EPQPAOPEPF	ARAPSPA_	PEPHPQP	PSPIPSP
104	TRIM47	miR-762(RP	298	CDS	-123	90,62	22	gcucgggcc	cugcagcuccg	LQLRQGSGPC	ARAPGPA_	LGPRVRP	SGPGSGP
105	TRIM65	miR-762(RP	310	CDS	-123	90,62	22	gccccgauc	cccgcccggga	PARDPGPDPC	APIPAPA_	PRSRPRP	PDPGPGP
106	USP51	miR-762(RP	460	CDS	-123	90,62	22	gcccggccc	cccaccccgc	PTPASSPAPA	ARPPPPP_	PGPRPRR	PAPAPAA
107	YAE1D1	miR-762(RP	481	CDS	-123	90,62	22	gcccggcuc	cacuccccgcu	HSPLPRPAPC	ARLRAPG_	PGSGPQA	PAPGPRP
108	ZC3H18	miR-762(RP	2429	CDS	-123	90,62	22	ucucggucc	ccguguccuca	PCPQPALGPF	SRSPAPA_	LGPRPQP	SVPGPSP

Figure 2: Results of processing the genome and miR-762.

For alignment of the nucleotide sequences of miRNA with mRNA, we assume the existence of only one admission on miRNA (lack of complementary couple for the hydrogen bond) that allows us to consider binding sites of mRNA longer than miRNA by one nucleotide. In this case, the regular structure of the spiral is broken, and a bulge exists. The free energy of binding of miRNA with mRNA for such a structure is less than that in an alternative case. The program determines the free energy of hybridisation (ΔG, 100 kJ/mole) of miRNA with mRNA and the scheme of their interactions; it also calculates the relationship ΔG/ΔGm, the levels of reliability (p), and the mRNA areas where the site (5′UTR, CDS or 3′UTR) starting at the first nucleotide of the 5′UTR is located. ΔGm is equal to the free energy of binding of miRNA with a site in nucleotide sequence of miRNA completely complementary to it. The level of reliability (p) was defined on the basis of the value of ΔG and its standard deviation. The program outputs the scheme of the interaction of miRNA with mRNA; a site position in 5′UTR, CDS, or 3′UTR; and the free energy of interaction of miRNA with mRNA and its relative value from the maximum energy of binding of miRNA. In the program, the threshold value of this relationship is set, and this value prevents consideration of sites with weak free energy of binding. In this article, an algorithm for scanning of genes on the cluster is provided **(Figure 1).**

Results and Discussion:

The developed algorithm has advantages that are not present in known programs for prediction of binding sites of miRNA with mRNA. With using of the realized fragmented algorithm on the cluster platform www.ursa.kaznu.kz data for a human genome were processed about 2700 miRNAs. miR-762 has 108 sites in the CDS of mRNAs with ΔG/ΔGm values greater than or equal to 90% **(see Figure 2).** miR-762 has 19 binding sites on the mRNA of the *CDKN1C* gene, 7 binding sites on the mRNA of the *TAF4* gene, 6 binding sites on the mRNA of the *BSN* gene. The binding sites are highly conserved. Nucleotide sequences of the binding sites in CDS encode olygopeptides containing mainly alanine and proline depending on the reading frame **(see Figure1).** On a

BIOINFORMATION
Discovery at the interface of physical and biological sciences

Open access

cluster 24 nodes were involved, the fragmented algorithm broke input data at first into 24 parts, then 9 of them into 24 fragments as 15 nodes finished the operation before remaining ones.

The algorithm of processing took 31,5 times less time, than when the sequential algorithm of scanning which processes the same data.
In the literature, there are many data regarding the value of the free energy of hydrogen bonds between nucleotides in water solution **[3]**. However, there is high variability in the value of the free energy of this bond and it is difficult to give preference to certain data **[4]**, **[5]**. It is important to know the relative relationships of the free energy of hydrogen bonds between nucleotides, as they are necessary for the formation of secondary and tertiary structures in RNA. The analysis of the free energy of the hydrogen bond arising between nucleotides during intramolecular interactions of mRNA in the formation of its secondary structure showed that three bonds formed between the G and C nucleotides, two between A and U, and one between G and U and between A and S. The relationship of the free energy of the hydrogen bond in G-C and A-U pairs approximately corresponds to the relationship of the forces of their 3:2 interaction (0.188 nNewton and 0.125 nNewton, respectively) **[6]**. The value of the free energy of one hydrogen bond between nucleotides varies in the range from -0.7 to -1.6 kcal/mol **[7]**. In this algorithm, the free energy of the interaction of nucleotides resulting from hydrogen communication was considered as equal to 6.368 kJ/mol and 4.246 kJ/mol for G-C and A-U pairs, and 2.123 kJ/mol for G-U and A-S pairs, respectively.

The distance between nucleotides in G-C and A-U pairs is 1.03 nm, whereas that between nucleotides in the G-U pair is 1.02 nm, and that between nucleotides in the A-S pair is 1.04 nm **[8]**. Therefore, formation of hydrogen bonds between these pairs of nucleotides enables the two-chained structure of mRNA to have a spiral form similar to that of DNA. Such a structure of mRNA without hydrogen bonds is stabilised by stacking interactions between the nitrogenous bases **[9]**. The distance between nucleotides in purine-purine and pyrimidine-pyrimidine pairs significantly differs from 1.03 nanometers; the distances for A-A, G-A, and G-G are equal to 1.23 nm, 1.25 nm, and 1.25 nm, respectively. In pyrimidine-pyrimidine pairs, the distances between nucleotides also significantly differed from 1.03 nanometers; the distances between nucleotides for C-C, U-U, and U-C was equal to 0.85 nm, 0.81 nm, and 1.18 nm, respectively. Therefore, in such pairs, hydrogen bonds are not formed, and these pairs will disrupt the regular structure of two-chained miRNA using the mRNA-reducing ability of the RISC complex (RNA-induced silencing complex). Therefore, in the program, such couples of hydrogen bonds were not considered.

According to **Table 1** it is obvious that the fragmented algorithm in comparison with the parallelized algorithm gives advantage on handling speed on large volumes of data. In case of increasing quantity of nodes of a supercomputer data are distributed on smaller fragments and this increases the general speed of execution of the program. So, for example, when using 15 nodes of a supercomputer the general speed of execution of the program, when handling 10000 miRNAs, becomes 6,6 times higher, than when handling the same number of data by the parallelized algorithm.

The technology of the fragmented programming allows not only distributing data and functions between nodes, and, besides, promotes optimum work of a supercomputer in general. So there is an opportunity to redistribute loading between nodes on the program course if any nodes have finished the work before other nodes, and other nodes are still loaded. While during the work with the parallelized algorithm such opportunity needs to be programmed separately.

Conclusion:
In this study, we have shown using of the realized fragmented algorithm on the cluster platform www.ursa.kaznu.kz data for seach binding sites of miR-762 in mRNAs of human genes. MiR-762 has 108 sites in the CDS of mRNAs with ΔG/ΔGm values greater than or equal to 90%. The feature of this miRNA is the presence of multiple binding sites in the mRNA of its target genes. miR-762 was high in both breast cancer cell lines and specimens, and its overexpression increased breast cancer cell proliferation and invasion **[10]**. miR-762 is involved in the matrix mineralization in mature osteoblasts, vascular smooth muscle cell calcification and other processes **[11-13]**. miR-762 is identified as a potential biomarker for monitoring the state of immunosuppression, allograft nerve **[14]**. Therfore study the characteristics of miR-762 binding sites is important for undestanding its functions.

References:
[1] Ivashchenko A. *et al. Bioinformation*. 2016 **12** (1): 15 [PMID: 27212839]
[2] Eric Londina *et al. Proc Natl Acad Sci USA*. 2015 **112:** E1106 [PMID: 25713380]
[3] Guckian K. *et al. J.Am.Chem.Soc.* 1996 **118**: 8182 [PMID:]
[4] Turner D. *et al. J. Am. Chem. Soc.* 1987 **109**: 3783
[5] Sugimoto N. *et al. Biochemistry.* (1987) **14**: 4554
[6] Boland T. & Ratner B. *Proc. Natl. Acad. Sci. USA* 1995 **92**: 5297
[7] Kool E. *Annu. Rev. Biophys. Biomol. Struct.* 2001 **30**: 1
[8] Leontis N. *et al. Nucleic Acids Res.* 2002 **30**: 3497
[9] Richard A., *et al. Biophysical Journal.* 1995 **69**: 1528-1535
[10] Li Y. *et al. Cell Prolif.* 2015 **48**: 643
[11] Gao F. *et al. Int J Biol Sci.* 2015 **11**: 109
[12] Gui T. *et al. Lab Invest.* 2012 **92**: 1250
[13] Mo X. *et al. Intractable Rare Dis Res.*2014 **3**: 12
[14] Rau C. *et al. J Biomed Sci.* 2013 **20**: 64

Edited by P Kangueane
Citation: **Ivashchenko *et al.*** Bioinformation 12(4): 237-240 (2016)