ARTICLES

Chinese Science Bulletin 2006 Vol. 51 No. 15 1836—1847

DOI: 10.1007/s11434-006-2054-8

A novel genome-wide full-length kinesin prediction analysis reveals additional mammalian kinesins

XUE Yu[1*], LIU Dan[1*], FU Chuanhai[1], DOU Zhen[1], ZHOU Qing[1, 2] & YAO Xuebiao[1]

1. Laboratory of Cellular Dynamics, University of Science & Technology of China/Hefei National Laboratory, Hefei 230027, China;
2. College of Agriculture, Zhejiang University, Hangzhou 310003, China

Correspondence should be addressed to Yao Xuebiao (email: yaoxb@ustc.edu.cn)

Received November 30, 2005; accepted June 2, 2006

Abstract Kinesin superfamily of microtubule-based motor orchestrates a variety of cellular processes. Recent availability of mammalian genomes has enabled analyses of kinesins on the whole genome. Here we present a novel full-length kinesin prediction program (FKPP) for mammalian kinesin gene discovery based on a comparative genomics approach. Contrary to previous predictions of 94 kinesins, we identify a total of 134 potentially kinesin genes from mammalian genomes, including 45 from mouse, 45 from rat and 44 from human. In addition, FKPP synthesizes 25 potentially full-length mammalian kinesins based on the partial sequences in the database. Surprisingly, FKPP reveals that full-length human CENP-E contains 2701 aa rather than 2663 aa in the database. Experimentation using sequence specific antibody and cDNA sequencing of human CENP-E validates the accuracy of FKPP. Given the remarkable computing efficiency and accuracy of FKPP, we reclassify the mammalian kinesin superfamily. Since current databases contain many incomplete sequences, FKPP may provide a novel approach for molecular delineation of kinesins and other protein families.

Keywords: kinesin, comparative genomics, CENP-E, full-length kinesin prediction program, FKPP.

Kinesins are microtubule-based motor proteins that perform diverse functions[1−6], including the translocation of vesicles, organelles, chromosomes, protein complexes, RNA-binding proteins (RNPs), etc. They also help to orchestrate microtubule dynamics and determine the morphology of cells[7−10]. Kinesins contain three functional regions: the motor domain, neck, and stalk[11]. The motor region contains amino acid sequences, highly conserved among the eukaryotic phyla, which are composed of a Walker A ATP binding motif and a microtubule-binding domain[11]. Outside the motor domain, kinesins show great sequence diversity, which led to the hypothesis that neck and stalk regions of kinesin specify cargo binding. Recent studies have indeed demonstrated that several kinesins attach to specific cargoes through interactions with adaptor proteins bound to these regions[11]. Based on location of the motor domain, kinesins can be classified in three categories: N-type, I-type, and C-type[11]. Despite great progress in the delineation of kinesin function, it remains unclear as to the total number of kinesin genes in mammalian genomes and their functional specificities.

Previously, most studies have been focused on functional identification of kinesins' ATP-binding motifs (~160 aa) in mouse proteome[12−14]. However, an integrated kinesin motor domain contains about 300 aa, including an ATP-binding motif and a nearby microtubule-binding motif. Thus, only studies of kinesins' ATP-binding motifs will provide limited insights for understanding the functional conservation and specificity of kinesins in molecular level. Moreover, molecular delineation of kinesin function requires information of full-length sequence while experimental identification of full-length sequence is labor-intensive and often limited by the quality of cDNA library. In addition, no measure has been attempted to validate the identity and faithfulness of kinesin sequences in the public database. Thus, there is an urgent need to develop an efficient and accurate *in silico* approach for guiding experimental biologists. Once the novel and full-length kinesins have been predicted using such an *in silico* method, cDNA of any putative kinesin can be sequenced while its cellular function can be assessed by experimentation. In this regard, the *in silico* method has two principal tasks in assisting molecular delineation of the kinesin superfamily: (1) in validation of the current database to synthesize full-length sequence of known kinesins; and (2) in a search for novel mammalian kinesins.

Another important issue to be addressed is classifying the components of kinesin superfamily into sub-

* These authors contributed equally to this work.

classes. Early phylogenetic evolutionary analyses on kinesins were limited to human, mouse, fly, worm, and yeast[12−14]. And in these studies, only sequences of the ATP-binding motifs were adopted rather than full-length kinesin motor domains. Given the recent completion of the rat genome (Rat Genome Sequencing Project Consortium; http://www.hgsc.bcm.tmc.edu/projects/rat/), comparative genomics for mammalian kinesin superfamily has become feasible.

To this end, we have developed a novel computational assay named full-length kinesin prediction program (FKPP) based on comparative genomics approach. Given many proteins are fragments without full length in public databases, the FKPP program could also be a general method for prediction the full length sequences of these fragments. Based on the prediction results of FKPP, we have also employed the full-length kinesin motor domains for classification of the kinesin proteins. And the prediction result is helpful for further experimental manipulation.

1 Materials and methods

1.1 Clean redundant

The initial loose "criteria" are used for maximizing the likelihood for kinesin identification. However, we obtain many redundant sequences based on sequence alignment. Some turn out to be the same gene differing only by a few amino acids, and many of them are alternative splicing variants. We then evaluate all predicted kinesins by BLAT[15], a fast and accurate local aligner that can be used in un-translated and translated modes. We do not pursue on the mosaic genes and if the coordinates of two predicted kinesins are the same or overlapped, we view them as different alternatively splicing (AS) isoforms of the same gene family and choose the longest sequence for further analysis.

1.2 FKPP for full-length kinesin synthesis

After homology searching, some kinesin sequences remain incomplete. Thus we employ a novel comparative genomics approach, named full-length kinesin prediction program (FKPP), to predict the full-length sequence of fragment kinesins based on their evolutionary conservation. We apply this approach to synthesize full-length kinesins from human, mouse and rat genomes. The scheme is shown below:

(1) Based on our homology searching and redundant cleaning outcome, we retrieve the amino acid sequences of putatively orthologous kinesins from those three organisms, respectively. Multi-sequence alignment is then conducted using the ClustalW/X program with default parameters as described[16]. If any of the three sequences is short and can only be aligned to others sequences with partial region, or is exceeding beyond the N-terminal or C-terminal of others sequences, we follow step (2). However, if three sequences can be aligned to the extent of full-length, we then follow step (3). And if step (3) is completed, we end the computational cycle.

(2) For this step, we assume three putatively orthologous kinesins: HsKIF-A, MmKIF-A and RnKIF-A. HsKIF-A and MmKIF-A are much shorter than RnKIF-A, which has excessive sequence beyond N-terminal or C-terminal. We use RnKIF-A sequence to run BLASTP from NCBI website with default parameters against human and mouse proteomes, respectively, in non-redundant GenBank database (nr protein database). The hits with >70% identity are accepted for further analysis. Then we may get N and M homology fragments in human and mouse proteomes with RnKIF-A respectively. There are HsKIF-A, Hs-F_1, Hs-F_2, ⋯, Hs-F_{N-1} in human, and MmKIF-A, Mm-F_1, Mm-F_2, ⋯, Mm-F_{M-1} in mouse. These homology fragments can be discontinuous, with some of them possibly overlapping. We assemble the overlapping fragments into a single longer sequence, from N′ and M′ discontinuous fragments in human and mouse, respectively. We complement the gaps directly by the corresponding sequence region of RnKIF-A and generate "chimeric" sequences for human and mouse HsKIF-A′ and MmKIF-A′. If more homology fragments of HsKIF-A and MmKIF-A are found, we then directly use the complemented parts of RnKIF-A to generated "chimeric" sequences HsKIF-A′ and MmKIF-A′, and then localize HsKIF-A′ and MmKIF-A′ to their genomes by BLAT, respectively, in translated mode[15]. Although other similar tools for identifying potential exon/intron structure in pre-mRNA/protein can also be used, the BLAT tool is chosen here.

If the full length "chimeric" sequence cannot be localized on genome properly, or the complemented stretches cannot be localized on genome, or original gene structure is heavily disrupted, or BLAT alignment SCORE decrease significantly (>10%), we consider that our chimeric sequence to be improper and reject it. Obviously, there are some different sites and microindels[17] within complemented parts compared to genome translated content, so we correct these sites and delete the indels based on genome content to ensure the complemented parts of the sequence have 100% identity to the human genome content. If a STOP codon is found in our corrected sequence, it is also rejected. Otherwise, we keep the results and get HsKIF-A'' and MmKIF-A''. Then we shift to step (1).

(3) We check the alignable region of the sequences. If one KIF has more stretches than other two sequences, we try to complement such gaps by these stretches. Then the "chimeric" sequences are localized to their own genome by BLAT, and the rules to accept our complement are same as in step (2). Then we shift to step (1). For convenience of the experimentalists, the FKPP approach is so easy that could be implemented by hand.

1.3 Experimental validation of FKPP

Antibodies against a synthetic peptide (PYLQTKHIEKLFTANC; BACHEM Americas) derived from the 36 aa sequence uncovered in our FKPP, but missing from the published sequence[6], were raised in rabbits using standard protocol as described. Immunoprecipitation and western blotting were carried out as described[9].

To validate if the *in silico* CENP-E is also centromere-associated as is the human CENP-E in the database, we carried out immunofluorescence labeling to visualize tubulin, CENP-E, and DNA essentially as described[9]. Mouse antibody 177 was chosen as the antibody was originally used for the discovery of human CENP-E[6].

2 Results

2.1 Full-length kinesin prediction by FKPP

As the first step toward homology searching, we retrieve amino acid sequences of all known kinesins and kinesin-like molecules from various organisms including yeasts (*S. cerevisiae* and *S. pombe*), worm (*C. elegans*), fly (*D. melanogaster*), mouse (*M. musculus*), rat (*R. norvegicus*) and human (*H. sapiens*) from the public databases including NCBI GenBank, Swissprot, Kinesin Home Page (http://www.proweb.org/kinesin/), and Ensembl. Each known kinesin sequence is used to search for homologues in other organisms' "proteome" by standard BLASTP. Kinesin prediction across yeasts, worm and fly is carried out at the sequence identity greater than 30% while the *E*-value is less than e^{-10}. For prediction among mouse, rat and human genomes, we set two criteria: (1) In the pair-wised comparison, alignable sequence similarity should cover more than 80% of entire length of the shorter one; (2) The identity of aligned sequence in paired regions should be greater than 30%.

Despite the completion of several mammalian genomes, it remains elusive as to the respective number of kinesin genes in mice, rat and human genomes since genome project does not provide full-length cDNA sequences. The only source for partial collection of full-length kinesins is Kinesin Home Page, which lists a total of 94 kinesins with 36 in human, 47 in mouse and 11 in rat. Among these listed kinesins, about 20 kinesins only have partial amino acid (~160 aa) sequences.

Experimental identification of full-length sequence of all mammalian kinesins becomes an infeasible task given the large number of molecules and limitation of richness of kinesin molecules in individual cDNA libraries; however, *in silico* prediction may facilitate the identification of full-length kinesin with ease. Thus, we developed a novel *in silico* full-length kinesin prediction program (FKPP) to synthesize "full-length" mammalian kinesins. FKPP is a comparative genomics based approach, based on the fact that the human, mouse, and rat proteomes are much conserved with ~21% (1743/8148) indel (short stretch deletion or insertion) events within rodent protein-coding sequences, and small insertions and deletions of 1—10 bp in length occur at 5% of the point substitution rate[17]. Since kinesins are highly conserved between mouse and human[12−14], we reason that a comparative genomics-based approach is reliable as the gene structures of most of the kinesins are conserved evolutionarily.

In this work, we are able to "synthesize" 25 full-length kinesins based on the partial sequences in the database. For example, MmKIF16A (GI: 2443266) (kinesin-3) has only 160 aa even after homology searching. But its rat putatively orthologue (RnKIF16A) has a full-length of 4614 aa (GI: 34857644). Our FKPP analysis yields a full-length MmKIF16A of 4529 aa. Fig. 1 offers a scheme for the FKPP analysis using KIF16A as an example. The original MmKIF16A sequence is first aligned with rat KIF16A followed by a BLASTP search in mouse proteome using full-length RnKIF16A sequence. Three partial sequences identified as matches to RnKIF16A are shown in Fig. 1(b). We assemble several genes such as mKIAA1300 (GI: 28972710), unnamed protein (GI: 26325666) and

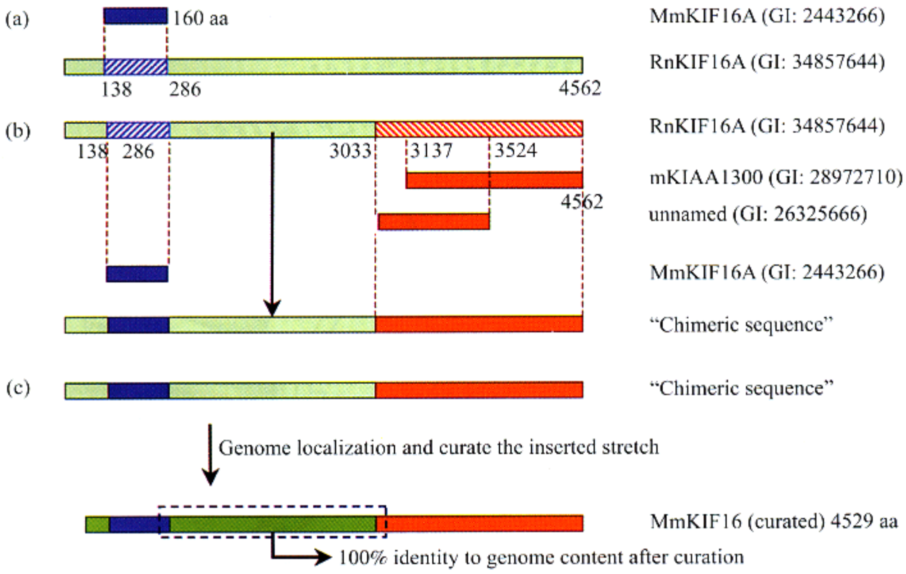


Fig. 1. FKPP for full-length kinesin discovery. (a) MmKIF16A is used as an example for FKPP given the availability of a short stretch of amino acid sequence available (160 aa). Partial sequence of MmKIF16A (GI: 2443266) is aligned with RnKIF16A. (b) Using RnKIF16A to conduct BLASTP search in mouse proteome, several homology sequences are found. For similarity, three sequences are presented MmKIf16A (GI: 2443266), mKIAA1300 (GI: 28972710), unnamed protein (GI: 26325666). The mKIAA1300 and unnamed proteins are first assembled into a unique sequence followed by filling the gap with sequence from the corresponding region of RnKIF16A, which leads to generation of a chimera. (c) The chimeric sequence to human genome with correction is done to make our sequence 100% identical to sequence published genome database. Then the *in silico* MmKIf16A contains a full-length of 4529 aa.

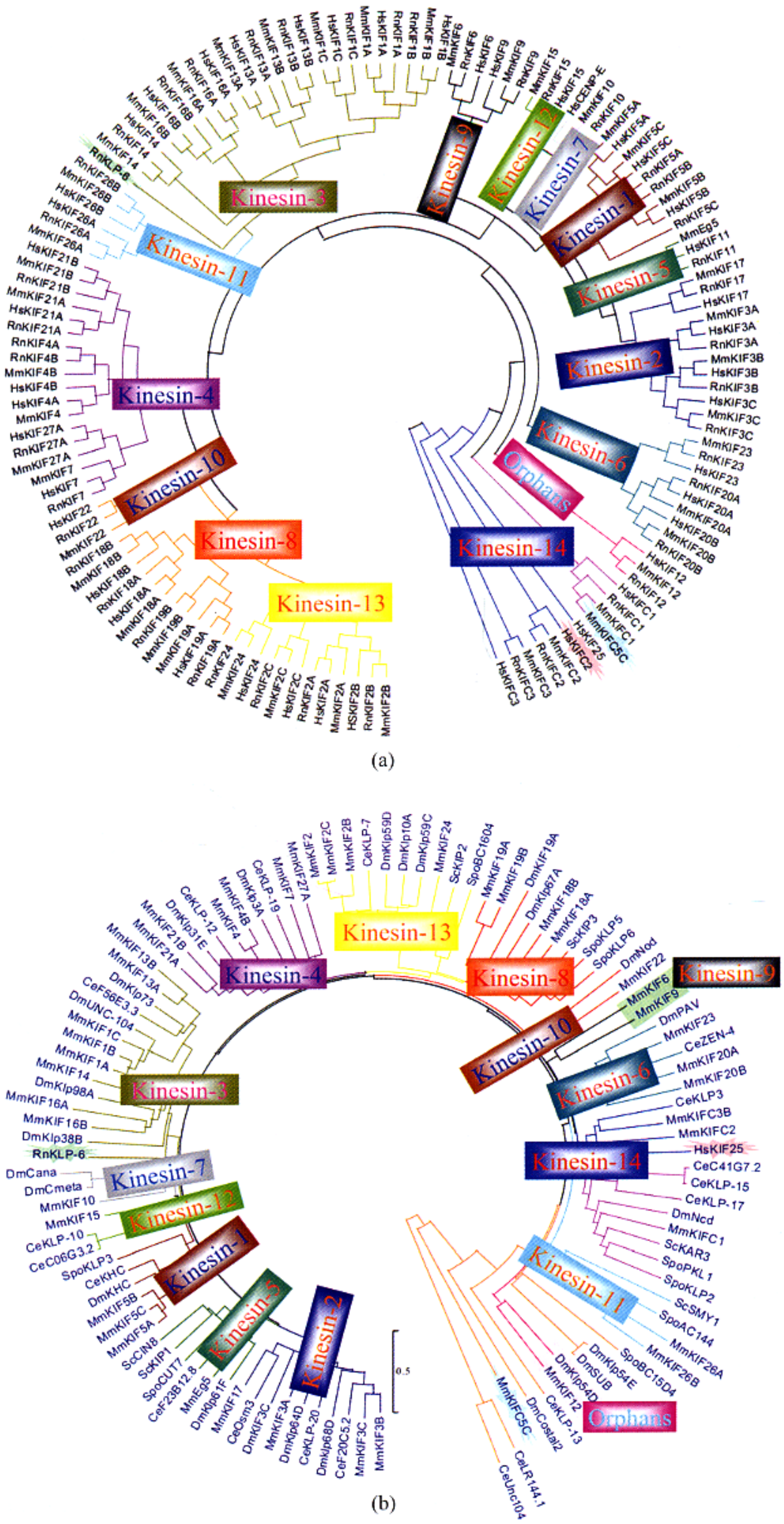
MmKIF16A (GI: 2443266) onto a unique template and directly fill the gap by the corresponding stretch in rat RnKIF16A, and generate a chimeric kinesin sequence. We then localize the chimera on the mouse genome by BLAT in translated mode. As shown in Fig. 1(c), modification is made to ensure chimeric sequence approaching a 100% identity to genome content.

2.2 *Phylogenic analyses of kinesin in mammalian and eukaryotic genomes*

After synthesis of all mammalian kinesins using FKPP, phylogenic procedures are carried out to classify the all kinesins into subfamilies across human, mouse and rat genomes. The sequences of various motor domains are aligned by ClustalW/X with manual curation. The phylogenic trees are then generated by the MEGA program (ver. 2.1) as previously described[18]. An evolutionary tree for mammalian kinesins is generated as shown in Fig. 2(a), by the Neighbor-Joining method with Bootstrap and the Poisson Correction. We also construct a phylogenetic tree for all seven organisms (budding yeast, fission yeast, nematode, fruit fly, mouse, rat and human), implemented in Minimum Evolution method with the Gamma Distance model (Fig. 2(b)). The bootstrap testing for the two phylogenetic trees have been performed to validate that our analyses are robust and reliable. All trees are un-root as previously reported[12−14].

Using FKPP, a total of 134 kinesins are identified including 45 from mouse, 45 from rat and 44 from human (Table 1). We adopt a standard kinesin nomenclature and prefixed with "Hs", "Mm", "Rn" for each kinesin to annotate organisms[19]. Comparison of these FKPP-synthesized with all known mammalian kinesins (Kinesin Home Page) led to an identification of 49 novel kinesins, including eight from human, 6 from mouse and 35 from rat. In addition, to re-classify the kinesins based on their functional domains and motifs, we also employ Interpro database[20] to analyze the novel kinesins uncovered by FKPP. And default parameters are chosen.

Our newly generated evolutionary trees are essentially consistent to the previous version[12−14]. Moreover, due to additionally kinesin discovered and used, our analyses provide more insightful information. Previously, KIF6, KIF7 and KIF9 were not classified into any kinesin sub-families and regarded to be orphan proteins[12−14]. However, in this work, both KIF6 and KIF9 have been classified into kinesin-9 sub-group. Interestingly, although the length of KIF6 and KIF9 are different largely, their motor domains are much similar, proposing that they may evolve from one ancestor. And



(a)

(b)

Fig. 2. Phylogenetic analyses with full-length kinesin motor domains. (a) Neighbor-Joining (NJ) tree with bootstrap for human, mouse, and rat. Three potential species-specific kinesins, HsKIF25, MmKIFC5C and RnKLP-6, are marked in the figure. (b) Minimum Evolution (ME) tree for mouse (human/rat), worm, fly, and yeast. We only use mouse kinesins as representatives for human, mouse and rat.

Table 1 A complete list of kinesins derived from FKPP in comparison with those previously published[a)]

Kinesin Home Page			After homology searching			FKPP
KIF	GI or accession	Length (aa)	Normalized name	GI or accession	Length (aa)	Length (aa)
HsCENP-E	399227	2665	HsCENP-E/HsKIF10	399227	2665	2701
HsKSP	1706622	1056	HsKIF11	1706622	1057	
KIF12	NP_612433	513	HsKIF12	32699596	551	618
HsRBKIN1	11761611	1805	HsKIF13A	21361722	1805	
HsGAKIN	8896164	1826	HsKIF13B	29421214	1835	
HsCMKrp	452517	1648	HsKIF14	23396633	1648	
HsKlp7	9910266	1388	HsKIF15	9910266	1388	
			HsKIF16A	34527855	323	4441
HsJ777L9	6522736	412	HsKIF16B	27529917	1393	1796
HsKIAA1405	7243191	791	HsKIF17	34978376	1029	
HsDKFZp434	12053149	898	HsKIF18A	21314742	898	
			HsKIF18B	37544008	870	
HsFLJ3730	NP_694941	548	HsKIF19A	23397458	548	894
HsATSV	2497523	1690	HsKIF1A	19924175	1690	
HsKIF1B	3043706	1338	HsKIF1B	42560524	1816	
HsKIF1C	3913961	1103	HsKIF1C	40254834	1103	
HsRabK6	3978240	890	HsKIF20A/MKLP2	5032013	890	
HsKlpMPP1	5911999	1780	HsKIF20B	15919888	1820	
HsNYREN62	5360129	633	HsKIF21A	33187651	1662	
HsLOC34389	XP_291594	512	HsKIF21B	41114119	1726	
HsKid	4519443	665	HsKIF22	6453818	665	
HsMKLP1	400264	960	HsKIF23/MKLP1	20143967	960	
			HsKIF24	34532133	850	1355
HsKlp6q27	4115553	384	HsKIF25	20138788	384	
HsKIAA1236	6330751	1481	HsKIF26A	20521808	1840	1887
			HsKIF26B	41114119	1726	
			HsKIF27A	30794488	1401	
HsKin2	3024057	679	HsKIF2A	4758644	679	
HsLOC8464	NP_115948	673	HSKIF2B	21707472	673	
HsMCAK	1695882	725	HsKIF2C	20141607	725	
HsKIF3A	3851492	702	HsKIF3A	33112673	702	
HsKIF3B	3913958	747	HsKIF3B	40788226	760	
HsKIF3C	3913957	793	HsKIF3C	3913957	793	
HsKIF4	7266951	1232	HsKIF4A	13959694	1232	
LOC347363	29743725	304	HsKIF4B	41147002	1234	
HsnKHC	2497520	1032	HsKIF5A	2497520	1032	
HsuKHC	417216	963	HsKIF5B	4758648	963	
HsxKHC	3043586	957	HsKIF5C	40788283	999	
			HsKIF6	ENSP00000287152[a]	482	525
			HsKIF7	38348350	830	1343
HsKIF9	11275982	725	HsKIF9	18202950	790	3308
HsCHO2	3702453	673	HsKIFC1	33875771	725	
			HsKIFC2	21955174	838	
HsKIFC3	12654739	694	HsKIFC3	34098691	694	
MmKIF10	2443268	160	MmCENP-E/MmKIF10	40388490	2474	
MmKIF11	2443270	170	MmKIF11	45476577	1052	
MmEg5	4160556	1014				
MmKIF12	12858387	642	MmKIF12	33563262	642	
MmKIF13A	10697238	1749	MmKIF13A	30794518	1749	1784
MmKIF13B	2443276	160	MmKIF13B	38076359	1960	
MmKIF14	2443278	166	MmKIF14	38073343	966	1662
MmKIF15	2443280	166	MmKIF15	38173736	1387	

(*To be continued on the next page*)

(*Continued*)

Kinesin Home Page			After homology searching			FKPP
KIF	GI or accession	Length (aa)	Normalized name	GI or accession	Length (aa)	Length (aa)
MmKIF16A	2443266	160	MmKIF16A	2443266	160	4529
MmKIF16B	2443262	150	MmKIF16B	38075140	2008	
MmKIF17	2443264	159	MmKIF17	23396634	1038	
MmKIF18A	12862603	151	MmKIF18A	21314852	886	
MmKIF18B	12862605	149	MmKIF18B	37537560	834	
MmKIF19A	12862606	148	MmKIF19A	12862607	148	1000
			MmKIF19B	38081372	748	
MmKIF1A	2506794	1695	MmKIF1A	2506794	1695	
MmKIF1B	2497524	1150				
MmKIF1Bbrain	5081553	1816	MmKIF1B	5081553	1816	
MmKIF1Bbeta	4512330	1770				
MmKIF1C	3913960	160	MmKIF1C	23821040	1100	
MmKlp174	1695174	887	MmKIF20A	6679597	887	
MmKIF20B	12862615	225	MmKIF20B	38085340	1774	
MmKIF21a	6561827	1573	MmKIF21A	6561827	1573	1638
MmKIF21b	6561829	1668	MmKIF21B	6561829	1668	
MmKIF22	2558833	148	MmKIF22	21704182	660	
MmKIFd19	12851512	457	MmKIF23	29568094	953	
MmKIF24	12862611	138	MmKIF24	45708952	1320	1327
MmKIFj19	12855902	157				
			MmKIF26A	38073760	1989	
			MmKIF26B	34328423	1550	1729
			MmKIF27	32401469	1394	
MmKIF2	125402	716	MmKIF2A	125402	716	
MmKIF2beta	2695866	659				
			MmKIF2B	38092011	706	
MmKIF2C	12862613	155	MmKIF2C	29840788	721	
MmKIF3A	125403	701	MmKIF3A	125403	701	
MmKIF3B	3122327	747	MmKIF3B	3122327	747	
MmKIF3C	3913959	796	MmKIF3C	3913959	796	
MmKIF4	1170659	1231	MmKIF4A	1170659	1231	
			MmKIF4B	20858575	1222	
MmKIF5a	3929108	1027	MmKIF5A	3929108	1027	
MmKIF5b	2062607	963	MmKIF5B	2062607	963	
MmKIF5c	3929110	956	MmKIF5C	3929110	956	
MmKIF6	2443284	165	MmKIF6	31581530	481	600
MmKIF7	2443286	168	MmKIF7	38086933	1328	
MmKIF8	2443288	185				
MmKIF9	5295882	790	MmKIF9	26325458	810	3205
MmKIFC2	1944330	792	MmKIFC2	1944330	792	
MmKIFC3a	2443294	157	MmKIFC3	12585614	709	
MmKIFC3b	12585614	709				
MmKIFC1	1944328	609				
MmKIFC5A	6979905	674	MmKIFC5A/MmKIFC1	13277705	674	
MmKIFC4	2558829	155				
RnKRP6	2674187	169	RnKIF11	34862680	1165	
			RnKIF12	34868461	617	
			RnKIF13A	34874048	1826	
			RnKIF13B	34874311	1903	
			RnKIF14	34880426	1659	
			RnKIF15	31335233	1385	
			RnKIF16A	62645648	4562	

(*To be continued on the next page*)

(*Continued*)

Kinesin Home Page			After homology searching			FKPP
KIF	GI or accession	Length (aa)	Normalized name	GI or accession	Length (aa)	Length (aa)
			RnKIF16B	34859296	1569	2254
			RnKIF17[b]			983
			RnKIF18A	34856640	659	804
			RnKIF18B	34873908	991	
			RnKIF19A	34875045	1046	
			RnKIF19B	34871266	1882	
			RnKIF1A	34877667	1943	
RnKIF1B	3493139	689	RnKIF1B	29789307	1816	
RnKIF1D	2370435	1097	RnKIF1C	22024392	1097	
			RnKIF20A	34878647	888	
			RnKIF20B	34862643	2017	
			RnKIF21A	34867881	1767	
			RnKIF21B	34880431	1670	
			RnKIF22	34859268	609	
			RnKIF23	34864667	896	947
			RnKIF24	34867277	1277	1320
			RnKIF26A	34935518	1933	
			RnKIF26B	34881041	1813	
RnKRP5	2674185	167	RnKIF27A	38016129	1394	
			RnKIF2A	34854206	771	
			RnKIF2B	27675158	712	
RnKrp2	2772516	671	RnKIF2C	20279134	671	
			RnKIF3A	34870729	708	
			RnKIF3B	34859022	562	747
RnKIF3C	3913949	796	RnKIF3C	16758244	796	
			RnKIF4A	34881081	1243	
			RnKIF4B	27668154	1224	
			RnKIF5A	34865745	1066	
			RnKIF5B	34876212	1114	
RnKHC	3122309	238	RnKIF5C	34854278	1004	
RnKRP3	2674181	160	RnKIF6	34874329	631	
			RnKIF7	34857299	1334	
			RnKIF9	34866378	3304	
RnKRP1	2674179	162	RnKIFC1	2674179	163	
RnKRP1	5070666	247	RnKIFC5A	34852223	616	
			RnKIFC2	38454244	791	
RnKRP4	2674183	153	RnKIFC3	34851230	739	
			RnKLP-6	34881054	1057	
			RnKIF10	34860597	2726	

a) There are a total 134 unique kinesins in our list, in comparison with 94 kinesins published previously. The nr database has been employed. Mouse and rat have 45 kinesins, while human has 44 KIFs, with HsKIF19B missing from our survey. b) The HsKIF6 sequence is retrieved from Ensembl database. c) The RnKIF17 is directly *in silico* elongated by mouse and human putatively orthologues.

KIF7 is grouped into kinesin-4 sub-family. Previous work identified a potential non-existed kinesin of KIF8 in mouse[12−14], which can be localized on its genome properly. Furthermore, this protein has no ortholog in either human and rat. This potential non-existed kinesin lead to obvious mistakes during classifying the kinesin N-2 sub-family. And in this work, we remove this protein to re-classify the sub-group. According to the standard nomenclature, a kinesin of KIF11 in original N-2 sub-group is classified into kinesin-5 sub-family. Previously, only KIF19A and KIF19B were identified and grouped into a sub-family together with KIF22. FKPP has generated additional two kinesins of KIF18A and KIF18B, which are highly similar with KIF19A and KIF19B. In this regard, we re-classify the components in this sub-group. The KIF22 is grouped into kinesin-10 sub-family, while KIF18A, B and KIF19A, B is classified into kinesin-8 sub-group. The detailed analyses on kinesin-8 sub-family are described below.

FKPP reveals that individual organism often contains species-specific kinesin such as MmKIFC5C, RnKLP-6, and HsKIF25. Therefore, it is of great interest to study their evolutionary traits. Based on phylogenetic analysis, we classify that MmKIFC5C and HsKIF25 into the kinesin-14 sub-family while RnKLP-6 belongs to the kinesin-3 subfamily. In addition, we propose that mammalian KIF24 is an I-type/M-subfamily (kinesin-13) member based on its short evolutionary distance from conventional M kinesins. The motor position of kinesin RnKIF24 starts ~170 aa downstream from the N-terminus, and is distinctly different from other typical N-type kinesins, which contain an additional ~50 aa after the motor domain to form the "neck" of the motor molecule. It would be of great interest to evaluate whether this kinesin moves and how it moves compared to other M kinesin without a neck region.

Despite vast information on classification of kinesin, previous work did not establish any link between mammalian KIF7 and other kinesin superfamily members. Based on our phylogenetic analysis, we assign mammalian KIF7 to kinesin-4 subfamily. In addition, we have identified a novel kinesin-4 subfamily member, mammalian KIF27, which contains 3 isoforms (A, B, C). Mammalian KIF24 and KIF25 are classified as members of N-11 sub-family (including KIF26A, KIF26B) in previous work[12−14]. However, our analysis shows that mammalian KIF24 is much like M KIF and therefore assigned to the kinesin-13 subfamily (including KIF2A, B, C and others). In addition, HsKIF25 is only in human as a species-specific KIF and assigned to the kinesin-14 subfamily. Although the overall sequences of mammalian KIF6 and KIF9 bear low homology, their sequences in the motor domain are almost identical, suggesting that they may have evolved from one ancestor. We thus assign them to the kinesin-9 subfamily.

Mammalian KIF18A, B, KIF19A, B (human only has KIF19sA) are very similar to fission yeast SpoKLP5 and SpoKLP6 and Budding yeast ScKIP3 in evolutionary distance, which all are classified in the kinesin-8 subfamily. ScKIP3, SpoKLP5 and SpoKLP6 were reported to be functional during mitosis and are localized on kinetochore[21]. Therefore, we propose that this subfamily may be localized to the kinetochore and involved in mitosis.

In order to validate our predicted kinesins to be real genes that can be expressed in human tissues, we also analyzed the expression profiles of human KIFs. We perform the homology search with 44 human kinesins in UniGene[22] and human ESTs databases, to testify whether these kinesin genes could be expressed in human tissue cells normally. For comparison, we also searched the GeneCards[23] database to identify the expression profile and tissue specificity of human kinesins. These results are listed in Table 2. Totally, 9 KIFs are identified as immune-specific for they are highly expressed in immune systems of >8－10 fold to other tissues. Since immune system is a specific organ with ubiquitous cells division, these immune-specific kinesins might play important roles during cell-cycle process. Surprisingly, at least four KIFs of them are which have roles in mitosis/cell division, KIF10/ CENP-E[9,10], KIF2C/MCAK[24], KIF11/EG5[25] and KIF20A/Rabkinesin-6/ MKLP2 (localized on midbody during cytokinesis, unpublished observation). Whether other five kinesins will be also functional during cell division should be experimentally verified.

2.3 Validation of FKPP by experimentation

Our early studies identified human CENP-E as a mitotic kinesin associated with the kinetochore[6]. Our recent studies demonstrate the importance of CENP-E as an essential motor for chromosome congression[9,10]. Surprisingly, FKPP predicts that full-length human CENP-E contains 2701 aa, 38 aa longer than that of published sequence[6]. In fact, recombinant full-length human CENP-E in insect cells was 5 kD shorter than that of the endogenous protein in human cells, suggesting that CENP-E cloned from the expression library may be somewhat incomplete. Fig. 3(a) displays *in silico* human CENP-E in relation to the published sequence[6]. To validate the accuracy of FKPP, we raised a peptide antibody against this fragment and used the antibody to isolate CENP-E from mitotic HeLa cells. As shown in Fig. 3(b), immunoprecipitates isolated by the newly made peptide antibody and a previously characterized CENP-E antibody HpX[7], but not rabbit IgG, contain CENP-E judged by the monoclonal antibody 177 that was used to clone human CENP-E[6]. Our triple immunofluorescence microscopic analyses indicate that both mouse CENP-E antibody 177 and the peptide CENP-E antibody gave indistinguishable labeling on the kinetochore of mitotic HeLa cells. Moreover, DNA sequencing of a human CENP-E clone isolated from the human testis library validates the accuracy of our FKPP analysis. Thus, human CENP-E con-

Table 2 The expression profiles analyses for human kinesins[a)]

Kinesin name	Accession ID	Length (aa)	Tissue-Specificity (GeneCard)	Expression profiles (UniGene/dbEST)
KIF1A	GI: 2497523[b)]	1690	neural specific (>8－10 fold)	neural, pancreas, bone marrow
KIF1B/KLP	O60333	1816	ubiquitous	ubiquitous
KIF1C	O43896	1103	ubiquitous	ubiquitous
KIF10/CENP-E	GI: 399227	2701	immune specific (>8－10 fold)	immune, liver, kidney, lung
KIF14	Q15058	1648	immune specific (>8－10 fold)	immune, liver, kidney
KIFC1	Q9BW19	673	immune specific (>8－10 fold)	immune, muscle, liver, pancreas
KIFC2	Q96AC6	838	neural specific (>5－10 fold)	neural
KIFC3	Q9BVG8	694	ubiquitous	ubiquitous
KIF3B	O15066	747	ubiquitous	ubiquitous
KIF3A	Q9Y496	702	ubiquitous	ubiquitous
KIF3C	O14782	793	neural specific (>8－10 fold)	neural specific (>8－10 fold)
KIF4A	O95239	1232	ubiquitous	ubiquitous
KIF4B	GI: 41147002	1234	/	testis (in mouse) /Hs.529460
KIF13A	Q9H1H9	1805	ubiquitous	ubiquitous
KIF13B	Q9NQT8	1826	ubiquitous	ubiquitous
KIF27A	Q86VH2	1401	ubiquitous	muscle, pancreas, kidney
KIF5A	Q12840	1032	neural specific (>8－10 fold)	brain, muscle, lung
KIF5B	P33176	963	/	neural specific(>8－10 fold)
KIF5C	O60282	957	neural and prostate specific (>8－10 fold)	ubiquitous
KIF17	Q9P2E2	1029	ubiquitous	spleen, brain
KIF11/EG5	P52732	1057	immune specific (>8－10 fold)	muscle, liver, lung
KIF9	Q9HAQ2	790	ubiquitous	ubiquitous
KIF22	Q14807	665	ubiquitous	ubiquitous
KIF25	Q9UIL4	384	ubiquitous	placenta, nervous /Hs.150013
KIF20A/MKLP2	O95235	890	immune specific (>8－10 fold)	ubiquitous
KIF2A/KIF2	O00139	679	Immune and neural specific (>5－8 fold)	Immune specific (>8－10 fold)
KIF2B	Q8N4N8	673	ubiquitous	medulla, testis /Hs.226805
KIF2C/MCAK	Q99661	725	immune specific (>8－10 fold)	immune, brain, muscle, liver
KIF23/MKLP1	Q02241	856	ubiquitous	ubiquitous
KIF18A	Q8NI77	898	immune specific (>8－10 fold)	testis, stomach /Hs.301052
KIF18B	37544008	870	immune specific (>8－10 fold)	ovarian, brain, bladder /Hs.406639
KIF12	Q96FN5	618	high in muscle, secretory, kidney (>5－8 fold)	ubiquitous
KIF20B	predicted[c)]	1897	immune specific (>8－10 fold)	testis, liver /Hs.240
KIF15/HKLP2	Q9NS87	1388		testis, liver /Hs.315051
KIF16A	GI: 41204881	846		
KIF16B	Q9HCI2	1393	ubiquitous	hippocampus, kidney, prostate /Hs.101774
KIF21A	AAR04774	1674		lung, kidney, liver /Hs.374201
KIF21B	GI: 41112866	1635	high in immune and nervous (>3－5 folds)	stomach, kidney /Hs.169182
KIF19A	Q8N1X8	894	ubiquitous	liver /Hs.372773
KIF26A	Q9ULI4	1840	ubiquitous	stomach /Hs.134970
KIF26B	GI: 41114119	1726	ubiquitous	immune specific (>8-10 fold) /Hs.125020
KIF24	predicted	1256		lung, liver /Hs.436169
KIF6	ENSP00000287152[d)]	484		
KIF7	predicted	1343		liver, stomach /Hs.528406

a) There are 44 kinesins in human analyzed. The tissue-specificity of gene expression profile is determined with GeneCard database. The electronic expression profile of kinesin genes (UniGene/dbEST) is taken from UniGene database with ESTs analysis. b) The kinesin is taken from GenBank database with GI number. c) The kinesin is predicted by FKPP. d) The kinesin is taken from Ensembl database.

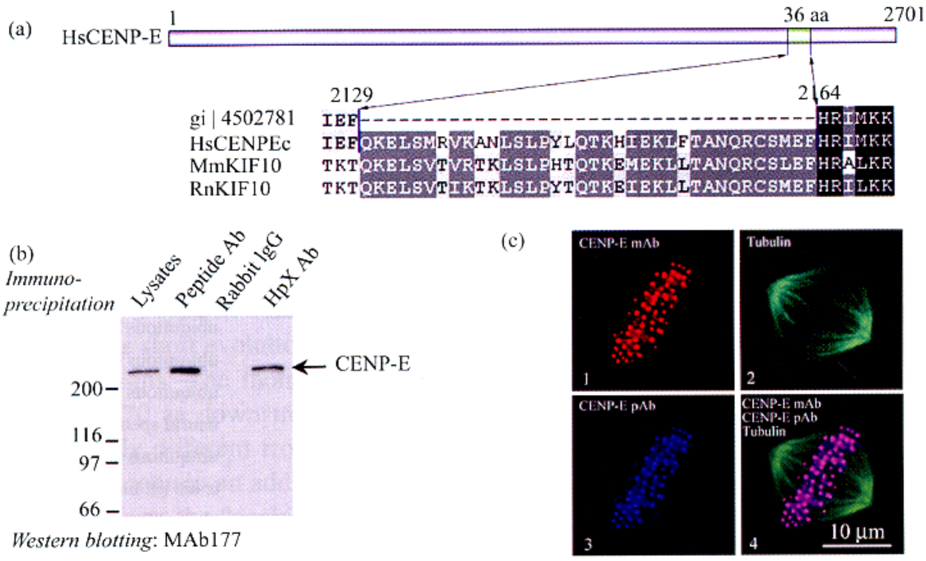


Fig. 3. FKPP reveals missing amino acids in human CENP-E. (a) Depiction of 36 aa revealed by FKPP but missed in human CENP-E of the current database. The corresponding amino acids for this missing segment are labeled (2119 and 2164). (b) Human CENP-E (FKPP-synthesized) represents the full-length human CENP-E. Mitotic HeLa cell lysates were prepared and incubated with protein A beads coupled with HpX antibody (HpX Ab), peptide antibody to the FKPP-derived CENP-E (Peptide Ab), and rabbit IgG, respectively. After washing, proteins bound to the antibody beads were resolved by SDS-PAGE and analyzed by western blotting using mouse antibody 177, the antibody employed to clone human CENP-E. (c) Human CENP-E synthesized *in silico* is a kinetochore-associated kinesin. 1, Confocal image of a mitotic HeLa cell labeled with CENP-E mAb177; 2, Confocal image of a mitotic HeLa cell labeled with a rat monoclonal antibody against tubulin (tubulin); 3, Confocal image of a mitotic HeLa cell labeled with CENP-E peptide antibody (CENP-E pAb); 4, Merge of 1—3.

tains 2701 aa. Validation of FKPP by three lines of ex--perimentation indicates that FKPP is a novel full-length kinesin prediction program with remarkable accuracy.

3 Discussion

Comparative genomics is a powerful tool for homologous gene prediction[26], whole-genome alignment and regulatory region prediction, in contrast to *ab initio* methods. Despite the fact that many organisms' genome-wide DNA sequences are available in the current database, the structural and functional relationship of individual genes remains to be established by wet-lab biologists. Many protein sequences still remain fragment status. We perform an estimation of the fragments in Swiss-Prot/TrEMBL. Approximately, we use key word of the name of the organism, such as "Homo sapiens" or "Mus musculus", etc, to find all protein sequences of the organism. And we use key word of the name of the organism plus "fragment" (i.e. "Homo sapiens fragment") to search all fragments in an organism. The result is shown in Table 3. There are about 32.2%, 24.6%, and 20.6% proteins which are fragments in human, mouse and rat, respectively. And even in budding yeast, there are still 2.8% of all proteins to be fragment.

Table 3 Fragment protein sequences of several organisms in Swiss-Prot & TrEMBL database.

Swiss-Prot & TrEMBL	Fragment	Total	Percentile
Homo sapiens	22779	70849	32.2%
Mus musculus	12136	49292	24.6%
Rattus norvegicus	2830	13716	20.6%
Xenopus laevis	2124	12182	17.4%
Arabidopsis thaliana	3017	43370	7.0%
Drosophila melanogaster	3119	28204	11.1%
Caenorhabditis elegans	352	22910	1.5%
Schizosaccharomyces pombe	273	5577	4.9%
Saccharomyces cerevisiae	414	14551	2.8%

Therefore, accurate *in silico* approaches are very helpful for guiding experimental biologists in high-throughput and high-content assays. Our FKPP analysis is a novel and easy method for maximal extraction of database information to facilitate the prediction and identification of novel mammalian kinesins. In addition, this method can be easily complemented with other gene prediction methods to improve the accuracy of gene prediction in eukaryotes including human. Due to the tools and database limitation, our prediction analysis may contain errors and inaccuracies and remains to be refined. For example, our analysis may miss some indel events, in which short stretches are present only in one organism but not others[17]. In addition, BLAT uses standard splice sites, so it may disrupt the gene structure if a given gene uses non-standard splice sites. However, no such case was found in our analysis for kinesins.

Our FKPP analyses support the notion that there are species-specific kinesins such as MmKIFC5C in mouse, RnKLP-6 in rat, and HsKIF25 in human as they do not have corresponding genes in other organisms. Do these species-specific kinesins contribute to the difference evolutionarily among the three organisms compared? This question requires further wet-lab characterization of these kinesins to show if they bear distinct biological functions.

Taken together, our analyses provide a foundation for future studies of the kinesin superfamily in cellular dynamics in mammals. Although the unique structural characteristics of each of kinesin subfamily suggest distinctly different mechanisms of motility and energetic considerations at the micro scale, molecular delineation of the specificity among kinesin subfamilies classified here will nevertheless provide a unified view of how kinesin motors work. In addition, FKPP provides a general and easy-to-use approach more than kinesins identification. Since many proteins are fragments at the current stage, FKPP is helpful to generating the potential full-length sequences among several similar organisms (i.e. human, mouse and rat). The prediction results are regarded as educated hypotheses before exquisitely experimental verification.

Acknowledgements This work was supported by the National Natural Science Foundation of China (Grant Nos. 39925018, 30121001, 30270293 & 90508002), the Chinese Academy of Sciences (Grant No. KSCX2-2-01), the Chinese 973 Project (Grant No. 2002CB713700), the Chinese 863 Project (Grant No. 2001AA215331), Chinese Minister of Education (Grant No. 20020358051), and American Cancer Society (Grant No. RPG-99-173-01).

References

1 Brendza R P, Serbus L R, Duffy J B, et al, A function for kinesin I in the posterior transport of oskar mRNA and Staufen protein. Science, 2000, 289(5487): 2120－2122

2 Cleveland D W, Mao Y, Sullivan K F. Centromeres and kinetochores: From epigenetics to mitotic checkpoint signaling. Cell, 2003, 112(4): 407－421

3 Guzik B W, Goldstein L S. Microtubule-dependent transport in neurons: Steps towards an understanding of regulation, function and dysfunction. Curr Opin Cell Biol, 2004, 16: 443－450

4 Hirokawa N, Takemura R. Molecular motors and mechanisms of directional transport in neurons. Nat Rev Neurosci, 2005, 6: 201－214

5 McIntosh J R, Grishchuk E L, West R R. Chromosome-microtubule interactions during mitosis. Annu Rev Cell Dev Biol, 2002, 18: 193－219

6 Yen T J, Li G, Schaar B T, et al. CENP-E is a putative kinetochore motor that accumulates just before mitosis. Nature, 1992, 359: 536－539

7 Abrieu A, Kahana J A, Wood K W, et al. CENP-E as an essential component of the mitotic checkpoint *in vitro*. Cell, 2000, 102: 817—826

8 Hirokawa N, Takemura R. Kinesin superfamily proteins and their various functions and dynamics. Exp Cell Res, 2004, 301: 50－59

9 Yao X, Abrieu A, Zheng Y, et al. CENP-E forms a link between attachment of spindle microtubules to kinetochores and the mitotic checkpoint. Nat Cell Biol, 2000, 2: 484－491

10 Yao X, Anderson K L, Cleveland D W. The microtubule-dependent motor centromere-associated protein E (CENP-E) is an integral component of kinetochore corona fibers that link centromeres to spindle microtubules. J Cell Biol, 1997, 139: 435－447

11 Vale R D, Fletterick R J. The design plan of kinesin motors. Annu Rev Cell Dev Biol, 1997, 13, 745－777

12 Miki H, Setou M, Hirokawa N. Kinesin superfamily proteins (KIFs) in the mouse transcriptome. Genome Res, 2003, 13: 1455－1465

13 Miki H, Setou M, Kaneshiro K, et al. All kinesin superfamily protein, KIF, genes in mouse and human. Proc Natl Acad Sci USA, 2001, 98: 7004－7011

14 Nakagawa T, Tanaka Y, Matsuoka E, et al. Identification and classification of 16 new kinesin superfamily (KIF) proteins in mouse genome. Proc Natl Acad Sci USA, 1997, 94: 9654－9659

15 Kent W J. BLAT: The BLAST-like alignment tool. Genome Res, 2002, 12: 656－664

16 Thompson J D, Gibson T J, Plewniak F, et al. The CLUSTAL_X windows interface: Flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res, 1997, 25: 4876－4882

17 Taylor M S, Ponting C P, Copley R R. Occurrence and consequences of coding sequence insertions and deletions in Mammalian genomes. Genome Res, 2004, 14: 555－566

18 Kumar S, Tamura K, Jakobsen I B, et al. MEGA2: Molecular evolutionary genetics analysis software. Bioinformatics, 2001, 17: 1244－1245

19 Lawrence C J, Dawe R K, Christie K R, et al. A standardized kinesin nomenclature. J Cell Biol, 2004, 167: 19－22

20 Mulder N J, Apweiler R, Attwood T K, et al. The InterPro Database, 2003 brings increased coverage and new features. Nucleic Acids Res, 2003, 31: 315－318

21 Garcia M A, Koonrugsa N, Toda T. Spindle- kinetochore attachment requires the combined action of Kin I-like Klp5/6 and Alp14/Dis1-MAPs in fission yeast. Embo J, 2002, 21: 6015－6024

22 Schuler G D. Pieces of the puzzle: Expressed sequence tags and the catalog of human genes. J Mol Med, 1997, 75: 694－698

23 Safran M, Solomon I, Shmueli O, et al. GeneCards 2002: Towards a complete, object-oriented, human gene compendium. Bioinformatics, 2002, 18: 1542－1543

24 Wordeman L, Mitchison T J. Identification and partial characterization of mitotic centromere-associated kinesin, a kinesin-related protein that associates with centromeres during mitosis. J Cell Biol, 1995, 128: 95－104

25 Blangy A, Lane H A, d'Herin P, et al. Phosphorylation by p34cdc2 regulates spindle association of human Eg5, a kinesin-related motor essential for bipolar spindle formation *in vivo*. Cell, 1995, 83: 1159－1169

26 Ureta-Vidal A, Ettwiller L, Birney E. Comparative genomics: genome-wide analysis in metazoan eukaryotes. Nat Rev Genet, 2003, 4: 251－262