interpreting predicted *Drosophila* genes.

Drosophila has been studied intensively and many of its genes have been well characterized experimentally. *Drosophila* protein sequences are collected in the SwissProt database and entries are updated periodically[2] to include information on alternatively spliced forms, alternative translation start sites, related motifs and functions of gene products.

We extracted 1,049 *Drosophila* protein sequences created before 1999 from the SwissProt database and compared them to the Celera proteome using BLAST[3] (Table 1). We found that 26.2% of the SwissProt sequences perfectly matched a sequence from the Celera *Drosophila* genome (CDG) and that 28.8% were of identical length, with at least 99% similarity over the length of the SwissProt protein. The remaining 45% had sequence differences of more than 1%, including mismatches, insertions and deletions — small and large — spread over the protein's length. We list four representative examples here (for other examples, see http://gnomic.stanford.edu).

The HMCU homeobox protein encoded by the *cut* gene is 2,175 amino acids long (SwissProt database), but there is no significant BLASTP match in the CDG proteome. The closest match is to a protein (CP31015) that contains a BTB homodimerization domain (average probability of error, $E = 10^{-18}$). Even much lower E values do not automatically imply homology, as BLASTP is strictly a local alignment algorithm that matches subsegments of query and target sequences across the whole protein database. Moreover, no normalizations are used to account for proteins with different lengths and quality[4]. For HMCU, only 14% of the complete SwissProt protein is matched.

The CIK2 voltage-gated potassium channel (encoded by the *shaker* gene) has a SwissProt length of 643 amino acids and a CDG length of 708 (92% identity with CP23511). There is no alignment for the first 49 residues of the SwissProt sequence and the first 90 of the CDG sequence. The corresponding DNA sequence is in the Celera genome, suggesting that the CDG prediction could have missed an exon.

The product of the *trithorax* gene is 3,759 amino acids long according to the SwissProt database, but has 3,085 residues according to CDG (81% identity with CP25007). At residue 238 of the SwissProt protein there is an 11-residue discrepancy due to a deletion of the dinucleotide TC in the Celera DNA sequence, followed by a GG insertion 33 bases later, which restores the translation frame. The CDG sequence has two large deletions corresponding to 27 and 34 amino acids, respectively, at positions 450 and 1,998, and a 12-amino-acid mismatch at position 238.

Table 1 Comparison of matching SwissProt and CDG proteins

Identity	*Same SP/CDG match length*	*SP/CDG match length within 1%*	*SP length < 99% CDG match*	*SP length >101% CDG match*
100%	276 (26.3%)	32 (3.1%)	43 (4.1%)	27 (2.6%)
>99%	302 (28.8%)	107 (10.2%)	56 (5.3%)	22 (2.1%)
<99%	2 (0.2%)	11 (1.0%)	78 (7.4%)	93 (8.9%)
Total	580 (55.3%)	150 (14.3%)	177 (16.9%)	142 (13.5%)

Many other examples of protein-sequence misalignments are available at http://gnomic.stanford.edu. The CDG contains two identical copies of the male-specific doublesex protein (CP29316 and CP39510). Other exact duplication pairs include CP24170/CP24220 and CP40382/CP40384. We compared the CDG proteins to all possible translations of the genome sequence and found in every case that there was only one match in the genome. The duplicate protein copies are probably annotation errors. SP, SwissProt.

The neuronal-differentiation protein Prospero has 1,403 residues according to SwissProt and 1,703 according to CDG (CP38193). The SwissProt protein does not contain the first 300 residues of the CDG sequence, but the remainder matches perfectly. The extra DNA in the Celera transcript matches the SwissProt gene (EMBL messenger RNA Z11743) upstream of the translation-initiation codon, suggesting that one of the sequences incorrectly predicts the translation start site.

There are numerous examples of differences between SwissProt and CDG sequences in the length of amino-acid runs — for example, a glycine run in HMOC (orthodenticle homeotic protein), an asparagine run in HMCA (homeotic caudal protein) and a glutamine run in CEB (enhancer-binding protein).

These and many other differences between the SwissProt and CDG sequences seem to arise predominantly from annotation mistakes, such as omission of alternative splice forms from the CDG predictions. Some could also result from sequencing and assembly errors. True polymorphisms in *Drosophila* are surely present, but their frequency and extent are unknown. Alleles are identified for some genes in Flybase[5], but many of these are deduced from targeted mutation experiments. The CDG annotation indicates that at least 1,600 genes disagree with their previously established sequences at the DNA level[1].

Gene-discovery algorithms exploit compositional differences between exons and introns, and look for signals such as splice sites and promoters. Searches using protein databases can help to locate genes that are similar to known genes but cannot identify completely new ones. Known protein-sequence properties could be used to improve exon prediction. Protein motifs and regular expression features such as zinc-finger motifs, ATP- and GTP-binding sequences, and motifs that are characteristic of kinase families are generally too small (5–15 residues) to be detected by BLAST, but may be used to support prediction of particular exons.

Another approach is to search for short 'words' (4–6 residues) that are significantly frequent[6]. Certain pentapeptides (such as GPPGP, CGKAF and HTGGK) occur frequently in human proteins but are very rare in non-coding sequences, making them suitable as coding-region indicators. Other unusual features such as clusters of charged or hydrophobic residues and high-scoring potential transmembrane segments could also prove useful[7]. Statistically significant charge clusters of basic or acidic residues occur in roughly 19–24% of higher eukaryotic proteins[8].

Although there are now more than 20 gene-prediction programs available, the task of annotation in eukaryotes is far from solved. Prediction solely on the basis of statistical and homology methods may prove to be intrinsically inadequate and experimental addenda may be needed. Results obtained by using expressed-sequence tags are valuable but tend to be biased by expression levels and are susceptible to contamination, fragmentation, fusion and other effects.

Proteomic studies using the present *Drosophila* genome sequence have significant limitations and the same will be true of the human genome. For now, these uncertainties should prescribe caution in interpreting newly predicted genes. Individual sequences should continually be corrected and refined by multiple rounds of annotation backed up with experimental data before the *Drosophila* genome can be considered complete and accurate.

Samuel Karlin, Aviv Bergman, Andrew J. Gentles

Mathematics Department, Stanford University, Serra Street, California 94305, USA

e-mail: karlin@math.stanford.edu

1. Adams, M. D. *et al. Science* **287,** 2185–2195 (2000).
2. Bairoch, A. & Apweiler, R. *Nucleic Acids Res.* **28,** 45–48 (2000).
3. Altschul, S. F. *et. al. Nucleic Acids Res.* **25,** 3389–3402 (1997).
4. Brocchieri, L. & Karlin, S. *J. Mol. Biol.* **276,** 249–264 (1998).
5. Flybase Consortium. *Nucleic Acids Res.* **27,** 85–88 (1999).
6. Karlin, S. & Cardon, L. R. *Annu. Rev. Microbiol.* **48,** 619–654 (1994).
7. Brendel, V. *et. al. Proc. Natl Acad. Sci. USA* **89,** 2002–2006 (1992).
8. Karlin, S. *Curr. Opin. Struct. Biol.* **5,** 360–371 (1995).

erratum

Flexible style that encourages outcrossing

Qing-Jun Li, Zai-Fu Xu, W. J. Kress, Yong-Mei Xia, Ling Zhang, Xiao-Bao Deng, Jiang-Yun Gao, Zhi-Lin Bai

Nature **410,** 432 (2001)

In the legend to Fig. 2, the full line refers to the cata-flexistyle flower and the dotted line to the hyperflexistyle flower (and not vice versa, as published).

Pollination

Flexible style that encourages outcrossing

Despite the convenience of self-pollination (selfing) in flowering plants[1–3], the detrimental effects of inbreeding that follow repeated selfing[3,4] have promoted strong natural selection for mating systems that ensure successful cross-fertilization (outcrossing). Here we describe a mechanism deployed by some tropical ginger flowers to avoid self-pollination — the flower moves its stigma (style), which normally acts as the pollen receptor, out of the way while its anther is releasing pollen. This cunning evasion adds to the diversity of pollination strategies that have contributed to the evolutionary success of flowering plants.

Alpinia is an Asian genus in the ginger family (Zingiberaceae) containing more than 250 species[5]. These are perennials with terminal inflorescences that produce between two and ten open flowers every day; each flower is hermaphrodite and lasts for only a day. We have monitored how the flower parts behave in nine species of *Alpinia*, both native and introduced, in a tropical seasonal rainforest in Xishuangbanna, Yunnan, in southwest China[6].

Each species of *Alpinia* has two phenotypes that coexist in all populations and which differ in the movement of the flower stigma (the phenotypes are termed cataflexistyled or hyperflexistyled flowers, depending on the direction of stigma movement during flowering). When cataflexistyled flowers are fully open (06:00–06:30), the stigma is held above the open (dehisced) anther from which pollen is being released (Fig. 1a). At the same time of day, the receptive stigma of hyperflexistyled flowers is curved downwards, below the indehiscent anther from which pollen has not yet been shed (Fig. 1b).

Flowers of both types retain these respective stigma positions until about midday, when the stigma of the hyperflexistyle form elongates and becomes erect above the anther (male phase). This movement prevents contact with insect visitors and creates an angle larger than 170° between the stigma and the anther's ventral face (11:45–13:30); the anther then dehisces and pollen is released (14:30–15:00; Fig. 1d).

The movement of the style of the cataflexistyle form is slower: here the stigma begins to move downwards and enter the receptive position (female phase; less than 170° from the anther's dorsal face) between 14:40 and 15:00 (several minutes after anther dehiscence in hyperflexistyle flowers; Fig. 1c). Flowering (or anthesis) ends in both forms after dark, when the anthers collapse and the corolla flops down.

The speed of stylar movement depends on the weather conditions, but all flowers of the same phenotype that open on the same day are strictly synchronous. The anthers of the hyperflexistyle flower never dehisce before all stigmas of the same phenotype have moved out of the receptive position (Fig. 2). It is likely that successful pollination only occurs between the two different forms, with the two phenotypes being associated with two genotypes (for example, in a natural population of *A. kwangsiensis* the ratio of individuals of the two phenotypes is about unity: 86:78; $\chi^2 = 0.39$, $P > 0.5$).

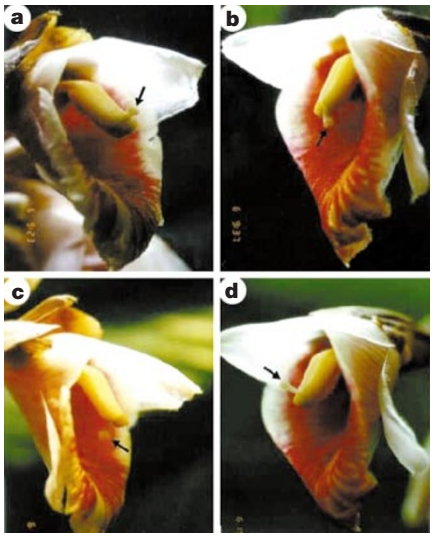

Figure 1 Positions of the stigma of the two flower forms in *Alpinia kwangsiensis* at different stages of flowering. **a,** Cataflexistyle flower in its male phase (before noon), in which the stigma is reflexed above the dehiscent anther. **b,** Hyperflexistyle flower in its female phase (before noon), in which the stigma is deflexed below the indehiscent anther. **c,** The same flower as in **a** during its female phase (afternoon), with the stigma below the anther; note that pollen has been removed from the anther by insect visitors (mainly xylocopid bees). **d,** The same flower as in **b**, but in its male phase (afternoon), with the stigma now erect above the anther, which then sheds its pollen. Arrows, stigma position.

We artificially manipulated different pollination combinations within and between phenotypes of *A. kwangsiensis* in the field. Our results indicate that fruit set resulting from cross-pollination between the two phenotypes is not significantly different ($F = 1.393$, d.f. = 1, $P = 0.242$) and that for the same treatments (self-pollination, cross-pollination, open pollination or controls), fruit-set rates did not differ significantly between the two phenotypes ($F = 2.251$, d.f. = 4, $P = 0.072$), indicating self-compatibility of the species. However, there was a significant difference between the treatments within the same phenotype ($F = 69.163$, d.f. = 6, $P < 0.001$): in both forms and during both gender phases, cross-pollination had a significantly higher fruit set than self-pollination, indicative of an inbreeding depression effect.

The floral strategy described here not only prevents self-pollination in a flower and within the same individual, but also among individuals of the same phenotype. It decreases inbreeding and promotes outcrossing in the plant by temporally and spatially separating the presentation of pollen and receptive stigmas through active floral movement. This mechanism, which we call flexistyly, differs from other passive outbreeding devices, such as dichogamy, herkogamy, enantiostyly and heterostyly[7], in that it combines some features of all of these mechanisms with the unique movement of floral parts.

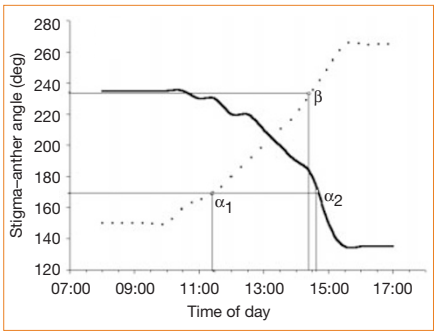

Figure 2 Floral behaviour of *Alpina kwangsiensis* during a single day of flowering. There is no overlap in the male and female phases of the two phenotypes (dotted line, cataflexistyle flower; full line, hyperflexistyle flower). α_1, Time when the stigma of a hyperflexistyle flower becomes reflexed out of its receptive position; α_2, time stigmatic receptivity of a cataflexistyle flower begins; β indicates the time of anther dehiscence of the hyperflexistyle flower.

We observed flexistyly in all nine *Alpina* species we studied[8]. In a molecular analysis of the phylogenetic relationships within the Zingiberaceae family (W. J. K. *et al.*, unpublished data), these nine species are distributed in three separate clades in the Alpineae, indicating that flexistyly either evolved independently several times in this Alpineae group or that it is widespread (though as yet unrecorded) in many taxa in the group (in *Amomum*, for example[9]).

Qing-Jun Li*‡, Zai-Fu Xu*, W. John Kress†, Yong-Mei Xia*, Ling Zhang*, Xiao-Bao Deng*, Jiang-Yun Gao*, Zhi-Lin Bai*

**Xishuangbanna Tropical Botanical Garden, The Chinese Academy of Sciences, Mengla, Yunnan 666303, China*
e-mail: qjlixtbg@bn.yn.cninfo.net
†*Department of Botany, National Museum of Natural History, Smithsonian Institution, Washington DC 20560-0166, USA*
‡*Kunming Institute of Botany, The Chinese Academy of Sciences, Kunming, Yunnan 650204, China*

1. Baker, H. G. *Evolution* **9,** 347–348 (1955).
2. Stebbins, G. L. *Am. Nat.* **91,** 337–354 (1957).
3. Darwin, C. *The Effects of Cross- and Self-fertilization in the Vegetable Kingdom* 2nd edn (Murray, London, 1916).
4. Holsinger, K. E. *Trends Ecol. Evol.* **6,** 307–308 (1991).
5. Smith, R. M. *Edinbr. J. Bot.* **47,** 1–75 (1990).
6. Zhang, J.-H. & Cao, M. *Biol. Conserv.* **73,** 229–238 (1995).
7. Richards, A. J. *Plant Breeding Systems* 2nd edn (Chapman & Hall, London,1997).
8. Li, Q.-J., Xu, Z.-F. & Xia, Y.-M. *Acta Botanica Sinica* (in the press).
9. Cui, X.-L., Wei, R.-C. & Huang, R.-F. in *Proc. 2nd Symp. Fam. Zingiberaceae* (eds Wu, T.-L., Wu, Q.-G. & Chen, Z.-Y.) 288–296 (Zhongshan Univ. Press, Guangzhou, China, 1996).