Conserved repeats in diverged ice nucleation structural genes from two species of *Pseudomonas*

Gareth Warren*, Loren Corotto and Paul Wolber

Advanced Genetic Sciences, Inc., 6701 San Pablo Avenue, Oakland, CA 94608, USA

Received 9 September 1986; Accepted 17 September 1986

ABSTRACT

Sequence analysis shows that an ice nucleation gene (inaW) from Pseudomonas fluorescens is related to the inaZ gene of Pseudomonas syringae. The two genes have diverged by many amino acid substitutions, and have effectively randomized the third bases of homologous codons. By reference to their potential for change, it is shown that certain conserved features must have been maintained by selection pressure. In particular, their conservation of internal sequence repetition, with three orders of repeat periodicity in each gene, suggests that the pattern of repetition is significant to the gene products' function. We propose models for the structure of the gene products in which each order of periodicity would be required for the nucleation function.

INTRODUCTION

Several genera of Gram-negative bacteria contain members able to nucleate the crystallization of ice in supercooled water (1,2,3). It is believed that nucleation is achieved by organizing large numbers of water molecules to form an ice crystal seed. The structures capable of doing this are of considerable interest, because they represent an unusual adaptation of biological molecules. In particular, these structures might provide insights into the interaction of biomolecules with water. Genes encoding ice nucleation activity have been cloned from Pseudomonas syringae (4) and P. fluorescens (5). The sequence of the inaZ gene from P. syringae was found to be internally repetitive (6), suggesting that a protein with repetitive primary and secondary structure may be responsible for ice nucleation.

A number of other proteins have been found to possess repetitive primary sequences (7). These include silk (8), the surface antigens of malarial circumsporozoites (9,10), Protein A from Staphylococcus aureus (11) and the antifreeze proteins of certain fish (12,13). The internal reiteration usually reflects the performance of similar functions by adjacent sections of the primary structure. Not surprisingly, most examples of internal repetition are found in proteins with non-enzymatic functions; it would be more difficult to understand the utility of possessing multiple sites for catalysis in cis.

A protein which organizes water molecules into an ice-like array must provide a template with very regular spacing, and thus it appears significant that the 24-base pair motif in inaZ is repeated with absolute regularity. How can its protein product be folded so as to present a suitable template? Certain additional features of the inaZ sequence may provide clues, particularly a six-fold periodicity superimposed on the basic motif. In this contribution we report that the inaW gene from P. fluorescens is a significantly diverged homologue of inaZ. The conservation of certain features of the periodicity appears to be significant, and holds important implications for the structure and function of the genes' products.

A DNA sequence encoding a repetitive protein may itself display unusual genetic properties as a consequence of phenomena related to recombination. Amplification and deletion of repeats can occur with relative ease. The resulting genes will often encode proteins which are still (at least to some extent) functional. Moreover, intragenic "gene conversion" may operate over an evolutionary time scale to make repeats more mutually similar than would be expected by selection for function. Thus the genetic plasticity of such DNA sequences adds to the difficulty of inferring functional significance from similarities in proteins from different species. On the other hand, the variation between repeats within a gene can provide additional information that is valuable in the understanding of function.

MATERIALS AND METHODS

Sequencing

A 5.4 kb DNA fragment, from the Kpnl site to the SalI site of pLVC46::kan111 (5), was cloned into vectors pUC18 and pUC19 (14). Both of the resulting subclones conferred ice nucleation activity equivalent to that conferred by previous plasmid clones. Mapping had indicated that the relevant gene(s) were confined to a region of 4.6 kb directly adjacent to the SalI site: a sequence determination was made for this region only. A range of restriction fragments was generated by partial digestion with Sau3A, HpaII, TaqI, or PstI, and complete digestion with EcoR1 or HindIII. These fragments were cloned into the M13 mp18 and mp19 vectors (14).

Single-stranded DNA from M13 clones was prepared from individual plaques on lawns of E. coli K12 strain JM101 (15), and sequenced by the dideoxy termination method (16) using an M13 universal primer (17). Two readings were made of each clone, and discrepancies corrected, before merging with overlapping sequences. The very exact sequence reiteration in one region (see below) made it possible to assign overlaps incorrectly. Therefore we determined Sau3A sites and PstI sites throughout the region by an end-labelling method (18), which yielded a high-resolution map against which the sequence information could be compared. The sequence is presented in Figure 1. From base 519 to base 4657, a region which contains the entire inaW gene (see below), all information is derived from readings of both strands. The remaining information has a lower level of confidence since it is mostly derived from readings of one strand. This was judged adequate for the analysis of any non-coding homologies.

Computer Algorithms

The principles of graphic matrix analysis of nucleic acid and protein sequences have been reviewed (19).

Alignment: The protein alignment function of program dFASTP (20) was used.

Aligned Comparison at each codon position: sliding windows of 20 codons from each gene were examined. The mean similarity at the first, second, and third positions of codons in the window was plotted in the upper, middle, and lower frames respectively of a tripartite graph.

Homology search: For each position x in sequence A and y in sequence B, a graph point was plotted at (x,y) if nucleotide N(x)=N(y), N(x+1)=N(y+1), N(x+3)=N(y+3) and N(x+4)=N(y+4). The omission of a test for (x+2,y+2) allows this algorithm to find a homology of two adjacent synonymous codons, even if their third positions have varied. A homology of significant length (coding or noncoding) would be displayed as a diagonal grouping of points. (Results of this analysis are described but not illustrated in the text.)

Octapeptide matrix comparison: Octapeptides were examined in the frame where the consensus octapeptide is YGSTLTAG. Lysine was considered identical to arginine, and glutamate to aspartate. For each position x in sequence A and y in sequence B, a large graph point was plotted at (x,y) if octapetide O(x)=O(y). A small graph point was plotted if O(x) differed from O(y) by exactly one residue.

Codon position-specific matrix comparison: Each sequence for analysis was separated into three sequences, which respectively received only the first, only the second, and only the third bases of the inaW and inaZ gene codons in the original sequence. For a position x,y in a sequence, a graph point was plotted only if nucleotide N(x)=N(y), N(x+1)=N(y+1), N(x+2)=N(y+2) and N(x+3)=N(y+3). To simplify the resulting graphs, not all values of x and y were examined: instead, the above test was made only where both x and y were integral multiples of four. This caused the algorithm to examine consecutive, non-overlapping tetranucleotide groups in the separated sequences, representing 4-codon groups in the analysis.

Criteria for graphic organisation of the translation products

Any hexadecapeptide which bears a more than 6/16 homology to

```
   1  atgaatgcataatgggcact ggatgacatagatggcgtac ggttgccatttatgggcccc aacgtgaccgaccatcgctt gggtattctttaattcccca gcggtaatctggaaggaaaa
 121  ctacttccgccaactcatcc ttctgtcagtgccacgtccg acagcaatgacctcGATCCC CTCTCGCGACGTCTCGCCTT TCTACTACCCCTTGCTTAAG CAGATTTATGCATAAAAGAG
 241  TATGGCGTTGACTCATCATT TTACGAGACAAGCCCAGTCA GAATTTATCGGCATACTGCG GACCGTCTACCCCCCAATGA CCGTTGACCTCATCcggctt ctcattacctattttgcatc
 361  cggcgactggatggtgctaa cagatcatcggcgtgcgcac ctggctgacacagattcacc aacanaccttagctttgctg tagttggaaatcagggcaaa atattctgacaaatgcgcca
 481  accgtagcttggttgagtta catgccagcaacttcatcTG ATGAGTATTGTTCTGCCCTC AAGATCGACGTATTTTTTGC AGTATCGCGCCATAATATTT ACTCGTTAGTTTGAGCATTA
 601  ATGGCTGCATAGACATCGTA TGTGAAAATGTATTATGGAC GATCCCCGTTAAAGGACTTA ATAGAAATGAGGCATAAGGG TAATTCAAAATGAAAAGTGA AAAAGTTCTGGTTTTACGGA
 721  CATGCGCTAATAATATGACC GACCATTGCGGTCTGGTATG GCCTATTTTAGGGCTTGTCG AATGCAAGTTTTGGGAGCCC ACAATAAAACTCGAAAATGG GTTGACCGGCGCACTGTGGG
 841  GACAAGGCTCAAGTGCCCAG TTGAGCATGAATGCAGACGC AAAATGGGTTGTTTGTGAAG TGACAATGGGCGACCTGATT TTCTTGGAAAATAATGAGGG GGTCAAGTTTCCTCGTGCAG
 961  AAGTAGTTCATGTCGGAACA CGAAGTAGCGCGCTAGGCTA CATTTCTGACAATGTTTCCA AGCATGAAGCATGTTCAAGT AATCTGATTGAGAAATTTAC TTTTTCTGATGTTAAATCAG
1081  AGACGAGGAATATTTCTCCC GCGCTGCCGGTTACTGTTGA CAACATGCCTAATGGCGTCA ATCGCAGCACAACTGTACGC AATACGCAAACGCTGGAAAC GGCCGTTTATGGCAGCACGC
1201  TCACTGGTGCTAATCAAAGC CAGCTCATTGCAGGCTACGG TAGCACCGAAACCGCTGGAG ATAGCAGCACTCTTATTGCG GGATACGGCAGTACCGGAAC TTCGGGCTCTGATAGTTCGA
1321  TCATTGCGGGTTACGGCAGC ACAGGCACCGCCGGCTCCGA TAGCTCACTCATTGCCGGAT ATGGCAGTACGCAGACTGCA GGCGGAGATAGCTCCCTAAC TGCCGGTTACGGTAGTACCC
1441  AAACGGCTCAGGTGGCAGT  AATCTCACCGCTGGCTATGG CAGCACCGGCACTGCGGGTC CTGACAGCTCACTTATCGCA GGTTATGGCAGCACGCAGAC TGCTGGGGGCGAAAGCTCCC
1561  TGACCGCCGGTTACGGCAGT ACCCAGACGGCTCAGGTGGG CAGTGATCTAACCGCTGGCT ATGGCAGCACCGGCACTGCA GGTTCCGATAGCTCGCTCAT TGCCGGATATGGCAGTACGC
1681  AGACTGCAGGCGGAGATAGC TCCCTAACTGCCGGTTACGG TAGTACCCAAACGGCTCAGG TGGGCAGTAATCTCACCGCT GGCTATGGCAGCACCGGCAC TGCGGGTCCTGACAGCTCAC
1801  TTATCGCAGGTTATGGCAGC ACGCAGACTGCTGGGGGCGA AAGCTCCCTGACCGCCGGTT ACGGCAGTACCCAGACGGCT CAGGTGGGCAGTGATCTAAC CGCTGGCTATGGCAGCACCG
1921  GCACTGCAGGTTCCGATAGC TCGCTCATTGCCGGATATGG CAGCACGCAGACTGCTGGGG GCGAAAGCTCCCTGACCGCC GGTTACGGCAGTACCCAGAC AGCTCAGGTGGGCAGTGATC
2041  TAACCGCTGGCTATGGCAGC ACCGGCACTGCAGGTTCCGA TAGCTCGCTCATTGCCGGAT ATGGCAGCACGCAGACTGCA GGTGGAGATAGCTCTCTGAC TGCCGGTTACGGTAGTACCC
2161  AAACGGCTCAGGTGGCAGT  GATCTCACCGCTGGCTATGG CAGCACCGGCACTGCAGGTT CCGATAGCTCGCTCATTGCC GGATATGGCAGCACGCAGAC TGCAGGCGGAGATAGCTCTC
2281  TGACTGCCGGTTACGGCAGT ACCCAAACGGCTCAGGTGGG CAGTGATCTCACCGCTGGCT ATGGCAGCACCGGCACTGCA GGTTCCGATAGCTCGCTCAT TGCCGGATATGGCAGTACGC
2401  AGACTGCAGGCGGAGATAGC TCCCTAACTGCCGGTTACGG CAGTACTCAAACGGCTCAGA TGGGCAGTAATCTCACCGCT GGCTATGGCAGCACCGGTAC TGCAGGTTCCGACAGCTCGC
2521  TCATTGCCGGATATGGCAGC ACGCAGACGGCAGGCGGAGA TAGCTCCCTGACCGCTGGTT ACGGCAGTACTCAAACTGCC GGTCACGGAAGCATCCTGAC TGCCGGATACGGCAGTACGC
2641  AGACGGCACAGGAAGGAAGT TCACTCACCGCAGGATATGG CAGCACGAGCACAGCCGGTC CTGAGAGTTCGCTGATCGCT GGCTACGGCAGCACCCAGAC TGCAGGACACGAGAGTACAC
2761  TCACTGCTGGCTACGGCAGC ACCCAAACAGCTCAGGAGGA TAGTTCACTTACTGCGGGAT ATGGCAGTACATCAACCGCA GGATTCAACAGCTCGTTAAT CGCAGGCTACGGTAGTACTC
2881  AAACCTCTGGGTATGAGAGC ATCCTGACCGCCGGCTACGG CAGTACGCAAACGGCCCAGG ATAATAGCTCTCTGACCACC GGCTACGGCAGTACATCGAC TGCGGGTTATCAAAGCTCGT
3001  TAATTGCGGGTTACGGCAGC ACCCAGACAGCAGGATATGA ATCTACGCTAACTGCAGGTT ACGGCAGTTGCCAGACGGCT CAAGAGCAAAGTTGGCTAAC TACCGGTTACGGGAGCACAT
3121  CAACTGCTGGCTACGAAAGC AGGCTGATTGCTGGATATGG CAGTACTCAAACCGCAGGCT ACAAAAGCATTCTAACCGCA GGCTACGGGAGCACTCAGAC CGCCCAGGAAGAGAGTTCGC
3241  TTACGGCCGGTTATGGCAGC ACTTCGACTGCGGGCTATGC AAGCTCCCTCATCGCTGGTT ATGGAAGCACCCAGACCGCA GGTTACGACAGTATCCTGAC CGCAGGTTATGGCAGCACGC
3361  TAACCGCTCTGGATAGCAGT ACGCTAACCGCAGGCTACGG TAGTACGGAAACAGCAGGAT TTGGCAGTTCGTTGATGGCC GGTTACGGCAGTTCGCAGAT CGCCGGTTATGGGAGCACGT
3481  TAACTGCGGGCTATGGTAGT ACCCAGATGGCAGAGCGGGA TAGCACTCTCACTGCTGGGT ATGGCAGCACCGGCACTGCG GGGCAGGACAGTTCCCTGAT TGCCGGTTATGGCAGCAGCT
3601  TGACCAGCGGCATGCGCAGC TATTTGACAGCCGGTTACGG CAGCACTTTAATCAGCGGAC TTCAGAGTGTATTAACGGCG GGGTACGGTAGCAGCCTTAC TTCAGGCATTCGCAGTAGCC
3721  TGACTGCGGGATACGGCAGC AACCAGATTGCGAGTCATAA GAGTTCTCTTATTGCGGGCC ATGAAAGCACTCAGATTGCA GGGCACAAAAGTATGTTGAT TGCCGGCAAGGGCAGTTCGC
3841  AAACAGCTGGTTCTCGCAGC ACGCTGATAGCTGGCGCTAA TAGCGTCCAGATGGCAGGGG ATCGTAGCAGGCTTACTGCC GGTGCAAACAGCATCCAAAC AGCGGGAGACCGCAGCAAGC
3961  TCCTGGCAGGCAGCAACAGT TATCTGACTGCTGGCGACCG GAGCAAACTCACCGCCGGCG ACGATTGTGTTCTGATGGCA GGGGATCGCAGCAAGCTGAC GGCGGGTAAGAACTGCGTAT
4081  TGACCGCCGGCGCTGACAGC AGACTCATAGGGAGCCTCGG CTCAACGCTCTCAGGTGGGG AGAACTCTACCCTGATTTTC CGATGCTGGGACGGCAAGCG TTACACCAATGTCGTCGTCA
4201  AGACCGGTACGGACGAGGTA GAGGCAGACGTTCCGTATCA AATTGACGAAGACAGTAATG TTCTAATTAAGGCAGAGGAC AATAGCGATACTCCGGTGGA CCAGTCTCAGATTCAGCCAT
4321  GAGACTTTTCTTTAAGCGAA GGAGGAGCCTAGGCTGAGCT TGCAGAGATGCTCTGCGGGT AAGGTGGCCGGGGTGGATGT TGACGCATCCTGACCGGCCG CTCCTTTTGGTTTCGAGGTA
4441  AGACTTGTTTTGCGCAGGCA CGGCGATCCCGCAAGGCATT CATGCCTTAGTCGGCGTAAC CATTAATTGAGTTGAATCGT AGAGTGTGCAGAGTTCAGAA GGCCCGTGTTCCGCAAGCCA
4561  GGCCGATGACGCAGAGAATG CTTAAGGCAAGCGAGTGGAA TGCTGAAGAGGCGCTGACGC GCCCCGCTAATCAGCACCAT TGAAACGTATTCAGATC
```

Figure 1. Nucleotide sequence of the DNA encoding inaW. One strand is presented in the 5' to 3' direction. Information derived from only one strand is presented in lower case letters. The following sequences are underlined: 649-664, extragenic homology with inaZ sequence (8/8, 10/12, or 12/16); 677-680, putative ribosome binding site; 690-692, inaW initiator codon; 4320-4322, inaW terminator codon; 4386-4433, putative transcriptional termination signal (three components: two 11/14 inverted repeats and a T-rich segment).

AGYGSTQTAGYGSSLT is presented as a separate word. Each group of three successive words is scanned for homology to AGYGST(G/S)TAGYGSSLI-AGYGSTQTAGYGSSLT-AGYGSTQTA(not-G)EGSNLT or a circular permutation thereof. If there is a match at four or more of the underlined positions, the sequence is presented according to rule I below; if not it is presented according to rule II. In both the InaW and InaZ proteins, blocks 2 and 3 fell under rule I, and blocks 1 and 4 under rule II. Rule I: a word is placed in the first, second, or third column according to which of the three hexadecapeptides above it matches most closely at the underlined positions. Rule II: successive words are placed up to three per line.

RESULTS AND DISCUSSION

Comparison of DNA sequences

The sequence containing the *inaW* gene of *P. fluorescens* was found to possess a single long open reading frame (Figure 1). This showed strong homology to the open reading frame of *inaZ* from *P. syringae* (6). However, only one apparently significant homology was found outside the respective reading frames. This occurred shortly 5' to the initiator codon in each case, and consisted of an 8/8 or 12/16 match (see Figure 1). It might have some significance to genetic regulation.

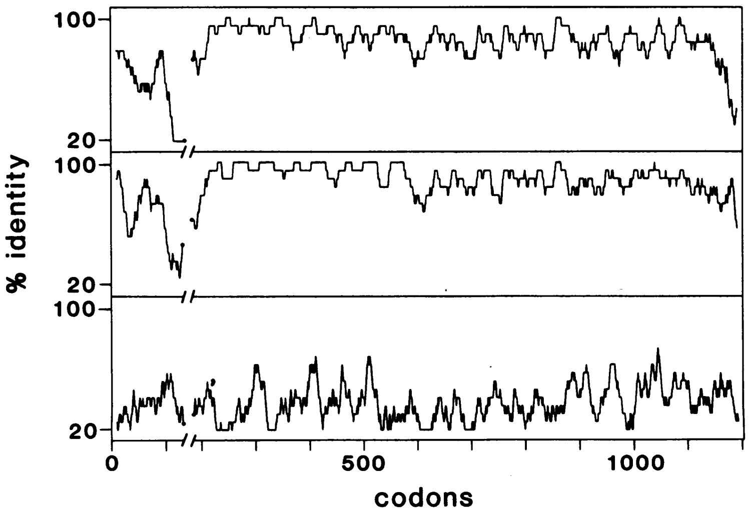

Figure 2. Comparison between aligned *inaW* and *inaZ* genes. Similarities between the first, second, and third positions of corresponding codons are separately plotted in the upper, middle, and lower graphs respectively. Mean values are plotted from a sliding window of 20 codons.

In order to draw inferences from the comparison of the InaW and InaZ proteins, it was necessary to know the extent to which their similarity had resulted from selection for maintenance of function. This was first estimated by comparing the similarity of aligned DNA sequences at each position of the codon. The _inaW_ and _inaZ_ sequences were aligned in two regions of strong homology, and the comparison is plotted in Figure 2. (Gaps in the alignment precluded the comparison of a single contiguous region). The relationship between corresponding third positions was observed to be effectively random, whereas nucleotides at the other two positions of the codon were strongly conserved in most regions. Therefore, it may be concluded that function has been conserved against mutational pressure, since the genes' divergence.

Comparison of repeat patterns

The translation products predicted for the _inaW_ and _inaZ_ genes are compared in Figure 3. Because the organisation of repeats is of key interest, the sequences are not merged; instead, each is divided into words, lines, and blocks by the same criteria. It is apparent that the criteria

```
      InaW                                                        InaZ
                                                                                                                Block
   1  MKsEKVLVLRTCANNMTDHCGLVWPILGLVECKFWEPTIKLENGLTGALW     1  MNLDKALVLRTCANNMADHCGLIWPASGTVESRYWQSTRRHENGLVGLLW
  51  GQGSsAQLsMNADAKWVVCEVTMGDLIFLENNEGVKFPRAEVVHVGTRss    51  GAGTsAFLsVHADARWIVCEVAVADIIsLEEPGMVKFPRAEVVHVGDRIs
 101  ALGYISDNVSKHEACSsNLIEKFTFSDVKSETRNISPALPVTVDNMPNGV   101  ASHFISARQADPASTSTSTLTPMPTAIPTPMPAVAsVTLPVAEQARHEVF
 151  NRsTTVRNTQTLET                                       151  DVASVsAAAAPVNTLPVTTPQNVQT

 165  AVYGsTLTGANQsQLI AGYGsTETAGDssTLI AGYGsTGTSGSDsSII   176                   ATYGsTLsGDNHsRLI AGYGsNETAGNHsDLI   1

 213  AGYGsTGTAGSDsSLI AGYGsTQTAGGDsSLT AGYGsTQTAQVGsNLT   208  AGYGsTGTAGSDsWLV AGYGsTQTAGGDsALT AGYGsTQTAREGsNLT   |
 261  AGYGsTGTAGPDsSLI AGYGsTQTAGGEsSLT AGYGsTQTAQVGsDLT   256  AGYGsTGTAGSDsSLI AGYGsTQTSGGDsSLT AGYGsTQTAQEGsNLT   |
 309  AGYGsTGTAGSDsSLI AGYGsTQTAGGDsSLT AGYGsTQTAQVGsNLT   304  AGYGsTGTAGSDsSLI AGYGsTQTSGGDsSLT AGYGsTQTAQEGsNLT   |
 357  AGYGsTGTAGPDsSLI AGYGsTQTAGGEsSLT AGYGsTQTAQVGsDLT   352  AGYGsTGTAGVDsSLI AGYGsTQTSGSDsALT AGYGsTQTAQEGsNLT   |
 405  AGYGsTGTAGSDsSLI AGYGsTQTAGGEsSLT AGYGsTQTAQVGsDLT   400  AGYGsTGTAGSDsSLI AGYGsTQTSGSDsSLT AGYGsTQTAQEGsILT   2
 453  AGYGsTGTAGSDsSLI AGYGsTQTAGGDsSLT AGYGsTQTAQVGsDLT   448  AGYGsTGTAGVDsSLI AGYGsTQTSGSDsALT AGYGsTQTAQEGsNLT   |
 501  AGYGsTGTAGSDsSLI AGYGsTQTAGGDsSLT AGYGsTQTAQVGsDLT   496  AGYGsTGTAGADsSLI AGYGsTQTSGSEsSLT AGYGsTQTAREGsTLT   |
 549  AGYGsTGTAGSDsSLI AGYGsTQTAGGDsSLT AGYGsTQTAQMGsNLT   544  AGYGsTGTAGADsSLI AGYGsTQTSGSEsSLT AGYGsTQTAQQGsVLT   |
 597  AGYGsTGTAGSDsSLI AGYGsTQTAGGDsSLT
                                                           592                                    SGYGsTQTAGAAsNLT   |
 629                   AGYGsTQTAGHGsILT AGYGsTQTAQEGsSLT   608  TGYGsTGTAGHEsFII AGYGsTQTAGHKsILT AGYGsTQTARDGsDLI   |
 661  AGYGsTSTAGPEsSLI AGYGsTQTAGHEsTLT AGYGsTQTAQEDsSLT   656  AGYGsTGTAGSGsSLI AGYGsTQTASYRsMLT AGYGsTQTAREHsDLV   |
 709  AGYGsTSTAGFNsSLI AGYGsTQTSGYEsILT AGYGsTQTAQDNsSLT   704  TGYGsTSTAGSNsSLI AGYGsTQTAGFKsILT AGYGsTQTAQERTsLV   |
 757  TGYGsTSTAGYQsSLI AGYGsTQTAGYEsTLT AGYGsCQTAQEQsWLT   752  AGYGsTSTAGYSsSLI AGYGsTQTAGYEsTLT AGYGsTQTAQENsSLT   3
 805  TGYGsTSTAGYEsRLI AGYGsTQTAGYKsILT AGYGsTQTAQEEsSLT   800  TGYGsTSTAGYSsSLI AGYGsTQTAGYEsTLT AGYGsTQTAQERsDLV   |
 853  AGYGsTSTAGYAsSLI AGYGsTQTAGYDsILT AGYGsTLTALDssTLT   848  TGYGsTSTAGYAsSLI AGYGsTQTAGYEsTLT AGYGsTQTAQENsSLT   |
 901  AGYGsTETAGFGsSLM AGYGsSQIAGYGsTLT AGYGsTQMAERDsTLT   896  TGYGsTSTAGFAsSLI SGYGsTQTAGYKsTLT AGYGsTQTAEYGsSLT   |
 949  AGYGsTGTAGQDsSLI                                     944  AGYGsTATAGQDsSLI                                     |

 965  AGYGssLTsGMRsYLT AGYGsTLIsGLQsVLT AGYGssLTSGIRssLT   960  AGYGsSLTSGIRsFLT AGYGsTLIAGLRsVLI AGYGssLTSGVRsTLT   |
1013  AGYGsNQIASHKsSLI AGHEsTQIAGHKsMLI AGKGsSQTAGSRsTLI  1008  AGYGsNQIASYGsSLI AGHEsIQVAGNKsMLI AGKGsSQTAGFRsTLI   4
1061  AGANsVQMAGDRsRLT AGANsIQTAGDRsKLL AGSNsYLTAGDRsKLT  1056  AGAGsVQLAGDRsRLI AGADsNQTAGDRsKLL AGNNsYLTAGDRsKLT   |
1109  AGDDCVLMAGDRsKLT AGKNCVLTAGADsRLI GsLGSTLSGGENSTLI  1104  GGHDCTLMAGDQsRLT AGKNsVLTAGARsKLI GsEGSTLSAGEDSILI   |

1157  FRCWDGKRYTNVVVKTGTDEVEADVPYQIDEDsNVLIKAEDNsDTPVDQS  1152  FRLWDGKRYRQLVARTGENGVEADIPYYVNEDDDIVDKPDEDDDWIEVK*
1207  QIQP*
```

Figure 3. Predicted amino acid sequences of the InaW and InaZ proteins, derived from sequences in Fig. 1 and ref. 6. The graphic presentation of each sequence is organised according to the same criteria. Amino acids are abbreviated as follows: ala, A; cys, C; asp, D; glu, E; phe, F; gly, G; his, H; ile, I; lys, K; leu, L; met, M; asn, N; pro, P; gln, Q; arg, R; ser, S; thr, T; val, V; trp, W; tyr, Y. Serines encoded by the AGPy-type codon are represented in lower case.

(which are arbitrary) similarly organise InaW-p and InaZ-p into four blocks of repeats, the blocks being of similar sizes in each case. To aid in the comparison, inter- and intra-genic homologies were plotted by an algorithm which examines successive, non-overlapping 8-residue groups (Figure 4). Their pattern indicated that repeats in Blocks 2, 3, and 4 of each gene are more similar to their counterparts in the other gene, than to other repeats present in cis. Thus, in spite of the potential for internal amplification to replace one block of repeats by another, the same block structure has persisted since the genes' divergence. Therefore this level of organisation is likely to be significant for function, even though we have recognised it only by adopting arbitrary criteria.

Within each protein, it is apparent that Block 2 shows the strongest self-homology, and Block 4 the weakest, with Block 3 intermediate. (Block 1 is too small for its self-homology to be meaningful.) However, when InaW is compared with InaZ protein, Block 2 does not show a greater homology to its counterpart than does Block 3. Therefore the more faithful repetition within Block 2 is probably due to more recent re-amplification or gene conversion activity, rather than implying that Block 2 has more stringent requirements for function.

A striking feature of each gene is that all octapeptides are contiguous, that a 16-residue periodicity is superimposed throughout, and that a further 48-residue periodicity is superimposed in Blocks 2 and 3.

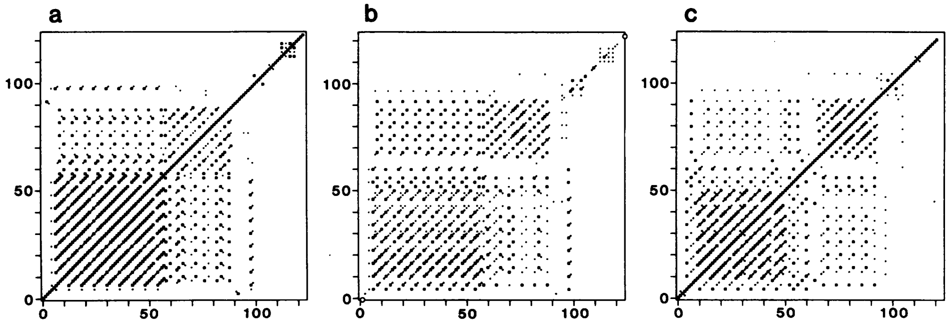

Figure 4. Comparison matrices of InaW against InaZ protein, and of each protein against itself. Small dots identify weaker octapeptide homologies than large dots. Parallel diagonal lines indicate repetition and their relative spacing represents the periodicity of the repetition. The criteria used in drawing these matrices are too stringent to detect homology between adjacent octapeptides. Weak diagonals are seen at 2-octapeptide spacing, and strong diagonals appear with 6-octapeptide spacing.

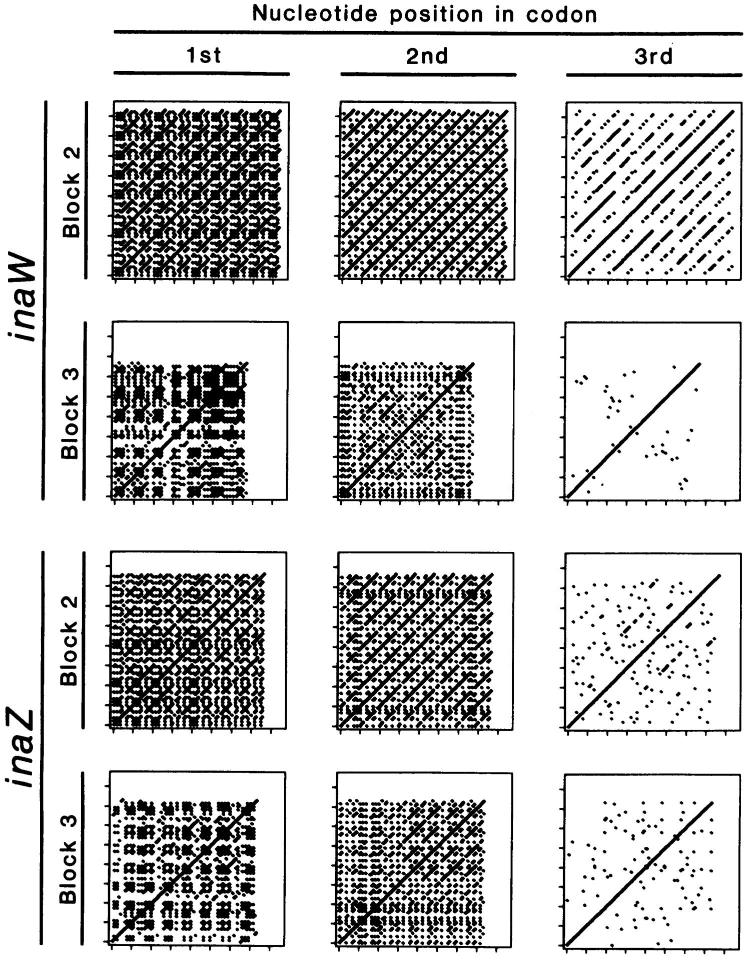

Figure 5. Self-homology matrices of Blocks 2 and 3 of inaW and inaZ, at each position of the codon. The sequences analysed were taken as follows: inaW Block 2, codons 213-628; inaW Block 3, codons 629-964; inaZ Block 2, codons 203-591; inaZ Block 3, codons 592-959. The graduations on the axes denote 48-codon intervals. Each dot represents exact homology between the appropriate positions in four consecutive codons.

Are these features necessarily maintained by selection? The alternative, again, is that amplification or gene conversion in cis is responsible. But unless the codon's third base shows periodicities similar to those in the first two bases, the other explanations break down and selection must be invoked. This was examined by an algorithm which plots separate self-comparison homology matrices for each position of the codon (Figure 5). In the third position, no significant periodicity was observed at the 8- and 16-codon level, which strongly implies that these levels of repetition are maintained by selection, and therefore have a functional role. At the 48-codon level, all codon positions in Block 2 of inaW show a very strong periodicity: thus amplification or correction has been active comparatively recently, and the 48-codon periodicity here need not be ascribed to selection. By contrast, the 48-codon periodicity in Block 2 of inaZ and in Block 3 of both genes was much weaker in the third position of the codon than in the first two positions. From this we infer that the 48-codon periodicity is selected, but that amplification or mismatch correction also play a role in its maintenance. The possibility of amplifying useful repeats and eliminating dysfunctional ones would reduce the mutational load of maintaining a functional gene.

The relationship between Blocks 2 and 3 in each protein is interesting. Block 2 ends and Block 3 begins at a different phase of the 48-residue periodicity in each case. However, the phase relationships are identical: in both cases Blocks 2 and 3 are out of 48-residue phase by +16 (or -32) residues. Thus there are two presumed independent occurrences of this phase relationship, whereas if only the 16-residue periodicity need be maintained, three possible phase relationships could be tolerated. It remains possible at present that the identical phase relationships arose by chance: we cannot estimate their functional significance from this comparison.

Fidelity of 8- and 16-residue repetition

It is natural to expect that the polynucleotide repeats encode similar functions. Given the (8x1) and (2x8) periodicities in the gene products, one might expect that all octapeptides share one function, but that alternating octapeptides are differentiated by the performance of different additional functions. This will be exemplified by the models described in the next section.

In order to ascribe constancy and alternation of function to particular positions of the octapeptide, we examined information from both inaW and inaZ. All 16-residue groups (a total of 123) were collated, and the results

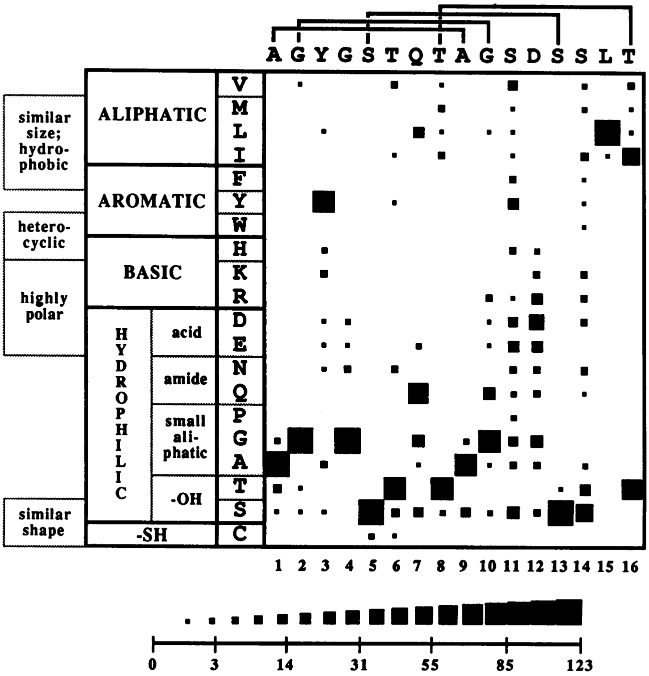

Figure 6. Patterns of substitution in the repetitive 16-residue word. Data were collected from all 123 words present in the InaW and InaZ proteins. The consensus word is shown at top; brackets above it link equivalent octapeptide positions which share the same consensus amino acid. The occurrence of each amino acid at each position is indicated by the area of the filled square at the corresponding coordinates (see scale, bottom). Amino acids are grouped into primary (solid boxes) and secondary (dashed boxes) classifications (21).

are shown in Figure 6. The eight positions in the octapeptide fall into two groups: those which show little or no alternation between successive octapeptides, and those positions at which neighbouring octapeptides differ more than 60% of the time. Among the former group, alanine and serine are very strongly conserved: they are presumably most important for the performance of whatever function(s) are shared by all octapeptides. At those positions which alternate, little similarity can be seen between the alternating residues.

Variability at any position of the 16-residue unit may be interpreted in one of two ways: either the various amino acids are acceptable substitutes in performing the same function, or the variation reflects a further differentiation of repeats between subtly different functions. At

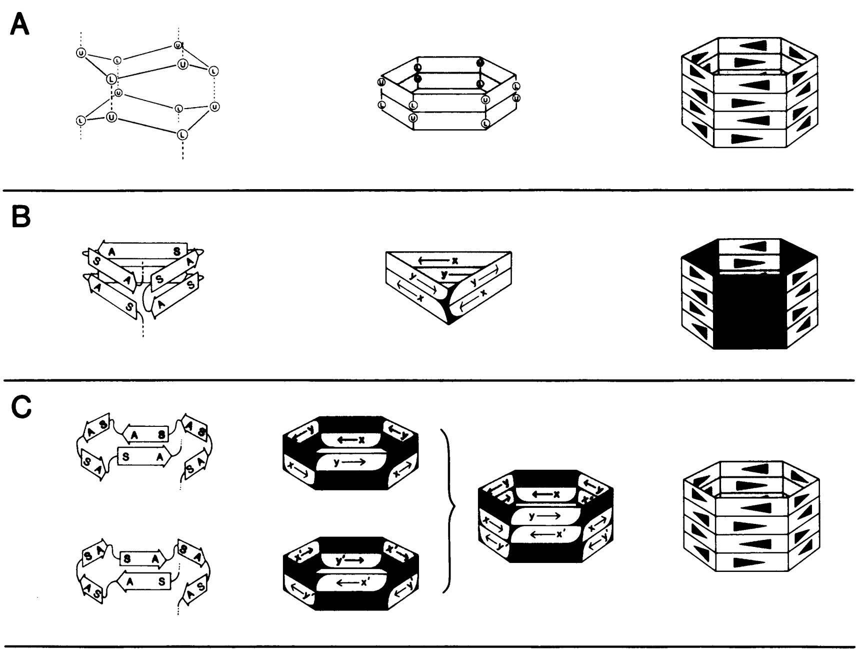

Figure 7. Structural models in which the Ina protein displays a symmetry related to that of ice. Drawings are not to scale. A: symmetry of ice I, illustrated by the spatial arrangement of 12 water molecules (at left), the symmetry of this structure (centre), and the way in which symmetry becomes extended as more water is added (at right). Lettered circles denote oxygen atoms; dashed lines represent O...H-O bonds in the c crystal axis. The labels U and L (upper and lower) serve for comparing the drawings at left and centre. B: triangular model; a 48-residue repeat is shown at left, and its symmetry is demonstrated at centre and right. S and A represent positions of serine and alanine residues respectively. The labels x and y represent an alternation of octapeptide types ((2x8) periodicity). C: antiparallel double helix model; a 48-residue repeat of each of the intertwined chains is shown at left. Their individual, combined, and extended symmetries are demonstrated at centre and right. S, A, x, and y are as above.

positions 7, 10, and 16, the latter explanation is favoured since the pattern of substitution is recognizable - it repeats after 48 residues. However, we cannot recognise any pattern in the substitutions at position 12, which include significantly large proportions of both acidic and basic residues. It may be surmised that strong polarity, regardless of net charge, is an important attribute at this position.

Constraints on protein modelling

The most striking feature of both ice nucleation genes is their precise periodicity, especially where three orders of periodicity overlap. Comparison of the genes has indicated that this feature has a functional significance. It is reasonable to assume that the products of these genes, in their native conformation, should possess a corresponding periodicity in their tertiary structure. This imposes a constraint on the structural modelling of such proteins. An additional constraint is suggested by the dogma that an ice nucleus functions by lattice-matching with ice. This leads to the expectation that the proteins' structure should be at least partially topotactic with the crystal structure of ice.

Even in the absence of direct evidence of the protein's structure, these two constraints restrict modelling considerably. Relatively few models can incorporate the necessity of all three orders of repetition (8x1, 2x8, and 3x16) in forming a structure topotactic with ice. Two such models

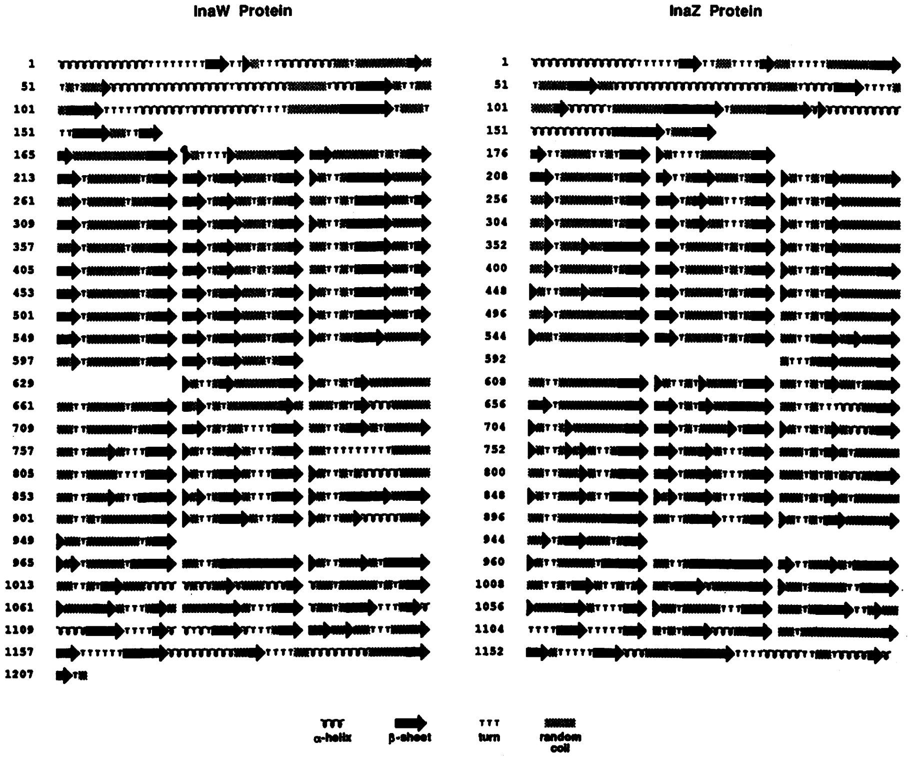

Figure 8. Predicted secondary structures (23) of the InaW and InaZ proteins. The layout of the predictions is isometric with that of the amino acid sequences (Figure 3).

are illustrated in Figure 7. The necessity for the (8x1) periodicity is explained by placing elements of each octapeptide in an invariant relationship to the nearest water molecules. Because alanine and serine residues are repeated with the greatest fidelity, we have used them to exemplify this relationship. The (2x8) and (3x16) periodicities are explained in different ways by the two models, although in both cases they are related to the "magic numbers" in the space group of ice I, P6-3/mmc (22). Both models make simple predictions: each octapeptide should be turned by at least 60^o relative to its covalent neighbour, and not all turns should be identical, although all should be performed by the equivalent portion of the octapeptide.

To refine these models, an algorithm (23) was used to make secondary structure predictions for both the InaW and InaZ proteins, and these are shown in Figure 8. They suggest that turns will be performed by the portion of the octapeptide whose consensus is GYG. They also suggest that much of the repetitive region will be beta-sheet. It is known that antiparallel stacking increases the stability of beta-sheets, whereas the predictive algorithm cannot take this into account. Within the postulates of models which predict extensive stacking (such as those in Figure 7), the algorithm's prediction of beta-sheet should be regarded as a particularly strong indication. The distances between the backbones of stacked beta-sheets are known from other proteins. The mean value for antiparallel stacking (24), 4.7 A, is not greatly different from the mean distance between planes in ice I (25), 3.7 A. We have not attempted to predict the other dimensions in our models: it will be more appropriate first to obtain direct experimental evidence of protein secondary structure.

ACKNOWLEDGEMENTS

We thank C. Dean, H.K. Dooner and T.H. Jukes for critical readings of the manuscript, and our colleagues for useful discussions and help. J.M. Thomas suggested the word "topotactic."

*To whom correspondence should be addressed

REFERENCES

1. Maki, L.R., Galyan, E.L., Chang-Chien, M.M. and Caldwell, D.R. (1974) Appl. Microbiol. 28, 456-459.
2. Schnell, R.C. and Vali, C. (1972) Nature 236, 163-165.
3. Lindow, S.E., Arny, D.C. and Upper, C.D. (1978) Phytopathology 68, 523-527.

4. Orser, C., Staskawicz, B.J., Panopoulos, N.J., Dahlbeck, D. and Lindow, S.E. (1985) J. Bacteriol. 164, 359-366.
5. Corotto, L.V., Wolber, P.K. and Warren, G.J. (1986) EMBO Journal 5, 231-236.
6. Green, R.L. and Warren, G.J. (1985) Nature 317, 645-648.
7. Ycas, M. (1972) J. Mol. Evolution 2, 17-27.
8. Lucas, F., Shaw, J.T.B. and Smith, S.G. (1957) Biochem. J. 66, 468.
9. Arnot, D.E., Barnwell, J.W., Tam, J.P., Nussenzweig, V., Nussenzweig, R.S. and Enea, V. (1985) Science 230, 815-818.
10. Godson, G.N., Ellis, J.; Svec, P., Schlesinger, D.H. and Nussenzweig, V. (1983) Nature 305, 29-33.
11. Uhlen, M., Guss, B., Nilsson, B., Gatenbeck, S., Philipson, L. and Lindberg, M. (1984) J. Biol. Chem. 259, 1695-1702.
12. De Vries, A.L. and Lin, Y. (1977) Biochim. Biophys. Acta 495, 388-392.
13. Lin, Y. and Gross, J.K. (1981) Proc. Natl. Adad. Sci. USA 78, 2825-2829.
14. Yanisch-Perron, C., Vieira, J. and Messing, J. (1985) Gene 33, 103-119.
15. Messing, J. (1983) Meth. Enzymol. 101, 20-78.
16. Sanger, F., Nicklen, S. and Coulson, A.R. (1977) Proc. Natl. Acad. Sci. USA 74, 5463-5467.
17. Sanger, F. and Coulson, A.R. (1978) FEBS Lett. 87, 107-110.
18. Smith, H.O. and Birnstiel, M.L. (1976) Nucleic Acids Res. 3, 2387-2395.
19. Maizel, J.V. and Lenk, R.P. (1981) Proc. Natl. Acad. Sci. USA 78, 7665-7669.
20. Lipman, D.J. and Pearson, W.R. (1985) Science 227, 1435-1441.
21. Dayhoff, M.O. (1969) Atlas of Protein Sequence and Structure. National Biomedical Research Foundation, Silver Springs, Maryland.
22. Burley, G. (1963) J. Chem. Phys. 38, 2807-2812.
23. Garnier, J., Osguthorpe, D.J. and Robson, B. (1978) J. Mol. Biol. 120, 97-120.
24. Dickerson, R.E. and Geis, I. (1969) The Structure and Action of Proteins, p35. Harper and Row, New York.
25. Eisenberg, D. and Kauzman, W. (1969) The Structure and Properties of Water, pp 71-79. Oxford University Press.