

# The internal oligopeptide sequences missing in crystals are disordererd domains

## Abstract

Polypeptide sequences in pdb format are invariably shorter than those in FASTA format. The missing residues are mostly internal oligopeptide strings and few C & N terminal residues. We have compared the panorama of the secondary structure domains generated from both formats by folding *in silico* and find that the missing oligopeptides are mostly from the intrinsically distorted domains.

**Keywords:** protein crystals, fasta format, pdb format, protein secondary structure, disordered domain,  $\alpha$  helix,  $\beta$  sheet, internal missing oligopeptides

## Research Article

Volume 7 Issue 1 - 2018

Natasha Kelkar,<sup>1</sup> Sohan P Modak<sup>2</sup>

<sup>1</sup>Institute of Bioinformatics and Biotechnology, S. P. Pune University, India

<sup>2</sup>Open Vision, India

**Correspondence:** Sohan P Modak, Open Vision, 759/75, Deccan Gymkhana, Pune 411004, India, Email [smodak@gmail.com](mailto:smodak@gmail.com)

**Received:** February 13, 2018 | **Published:** February 23, 2018

## Introduction

Prior to their maturation as a biological structure or function, nascent polypeptides fold to form three dimensional structures composed of  $\alpha$  helices,  $\beta$  sheets and disordered regions. The amino acid sequence of the processed polypeptide is stored in FASTA format ([www.rcsb.org](http://www.rcsb.org)) and it is almost always longer than that in the crystal structure, retrievable in PyMol stored in pdb format, wherein the absence of residues has been noted at the C-terminal, N-terminal and at intra-polypeptide locations of crystals. Indeed, a large number of protein crystals in the data base exhibit internal missing string.<sup>1</sup> Crystallographers generally consider that the missing residues are due to low electron density undetectable in low resolution crystallography. Since some of the gaps at the N and C termini can be attributed to post-translational processing, the presence of missing internal oligopeptides may lead to misinterpretation of the secondary structure domains in the immediate vicinity of the gaps as well as in the flanking segments. While studying the phylogeny of proteins<sup>2</sup> we considered the possibility that the extent of evolutionary conservation of residues defining individual secondary structure domains may be one of the determinants. As we came across the cases of internal missing intra-molecular residues here we analyze their structure and significance.

## Materials and methods

Amino acid sequences of 9 proteins were downloaded from RCSB pdb in FASTA and crystal format.<sup>3</sup> ([www.rcsb.org](http://www.rcsb.org)) These are, (1) SAICAR synthase from *Saccharomyces cerevisiae*, strain ATCC 204508/S288c (PDB Id: 1A48),<sup>4</sup> (2) SAICAR synthase complexed with ADP, AICAR, and succinate from the same strain as above ([www.rcsb.org](http://www.rcsb.org)), (3) Lipote-protein ligase A from *Streptococcus agalactiae* (PDB Id: 2P0L) ([www.rcsb.org](http://www.rcsb.org)), (4) P450 pyridoxylase from *Sphingopyxis macrogoltabida* (PDB Id : 3RWL),<sup>5</sup> (5) Hydroxymethylbilane synthase from *Escherichia coli* (K12) (PDB Id : 2YPN),<sup>6</sup> (6) UDP-n-acetylmuramoyl-L-alanine:D-glutamate ligase from *Escherichia coli* (K12) (PDB Id: 1UAG),<sup>7</sup> (7) Glycinamide ribonucleotide synthetase

from *Escherichia coli* (K12) (PDB Id: 1GSO)<sup>8</sup> (8) Polypolyglutamate synthetase from *Lactobacillus casei* (PDB Id : 1FGS)<sup>10</sup> and (9) mitochondrial helicase suv3 from *Homo sapiens* (PDB Id : 3RC3).<sup>11</sup>

The amino acid sequences in two formats were aligned and residues missing at the N-terminal, C- terminal and internal regions were detected. Sequences of 9 proteins were folded with JPred 4 (<http://www.compbio.dundee.ac.uk/jpred4>) and PSSPred.<sup>12-13</sup> From the output we designated residues forming secondary structure domains in different shades, namely light gray ( $\alpha$  helix), dark gray ( $\beta$  sheet/loop) and medium gray (disordered domain). The sequences derived from the crystals (pdb format) were similarly shaded.

## Results

The sequences in both formats of nine proteins are shown in Figure 1. We noticed that, in contrast to the sequence derived from FASTA file, some amino acids were missing at the termini as well as at internal locations of the polypeptide in the crystal-derived sequences. Upon folding these *in silico* with Jpred4, we find (Figure 1) that each polypeptide gave rise to lawns exhibiting  $\alpha$ -helix (○),  $\beta$  sheet (●), and disordered domains/random coil (●) (methods). Since the folding pattern with respect to the number and positions of different structural domains was nearly similar with PPSPred, we have restricted this presentation to JPred4 for proteins no 1-9.

Table 1 shows the number of amino acid residues missing in crystal-derived sequences. 2P0L, 3RWL and 3RC3 also exhibit long missing oligopeptides at the termini. Indeed, all crystal-derived sequences contain one or more 3-33 long internal missing oligo-peptides. Table 2 describes the distribution of missing residues in crystals based on their physicochemical properties and number. These were highlighted in sequences from mature protein (FASTA file) in Figure 1. We find that, in 10 cases, more hydrophilic residues are missing in the internal oligopeptide. In the rest 6, the ratio of hydrophobic residues to total number of missing residues is more than 0.5.

## (A) 1A48 (JPred4)

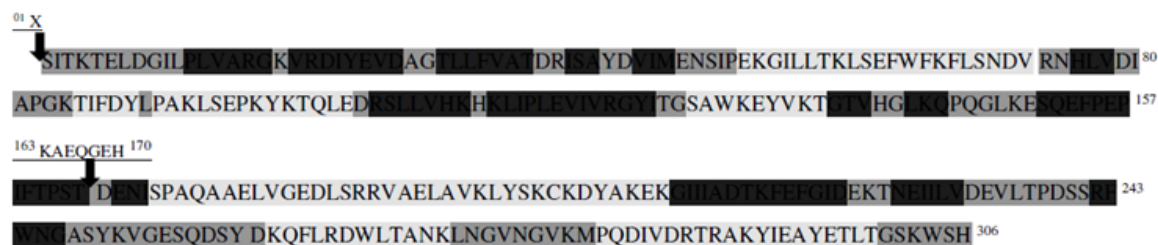
## Mature protein

SITKTELDGILPLVARGKVRDIYEVDAGTLFLVATDRISAYDVIMENSPEKGILLTKLSEFWFKFLSND  
 VRNHLVDIAPGKTIFDYLPKLSPEPKYKTQLEDRLVHKKHKLIPLEVIVRGYITGSAWKEYVKTGTVH  
 GLKQPQGLKESQEPPEPFTPTSTKAEQGEHEDENISPAQAELVGEDLSRRVAELAVKLYSKCKDYAKEK  
 GHADTKFEFGDEKTEHILVDEVLPDSSRFWNGASYKVGESQDSYDKQFLRDWL TANKLNGVNGV  
 KMPQDIVDRTRAKYIEAYETLTGSKWSH

## Mature protein folded with JPred4



## Crystal derived sequence



## Crystal derived sequence folded with JPred4



## (B) 2C9Q (Suicase complexed with ADP, AICAR, succinate)

## Mature protein

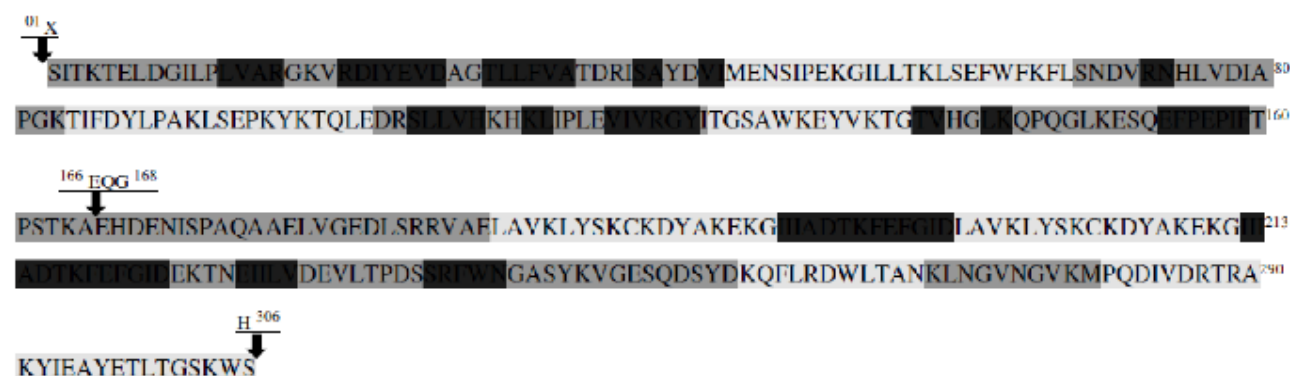
SITKTELDGILPLVARGKVRDIYEVDAGTLFLVATDRISAYDVIMENSPEKGILLTKLSEFWFKFLSND  
 VRNHLVDIAPGKTIFDYLPKLSPEPKYKTQLEDRLVHKKHKLIPLEVIVRGYITGSAWKEYVKTGTVH  
 GLKQPQGLKESQEPPEPFTPTSTKAEQGEHEDENISPAQAELVGEDLSRRVAELAVKLYSKCKDYAKEK  
 KMPQDIVDRTRAKYIEAYETLTGSKWSH

## Mature protein folded with JPred4

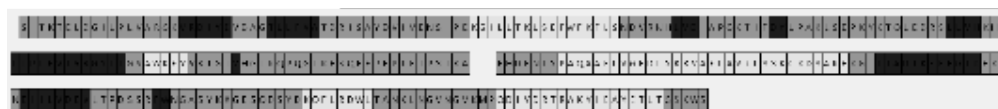


## Crystal derived sequence

(Figure 1 continues)



Crystal derived sequence file folded with JPred4

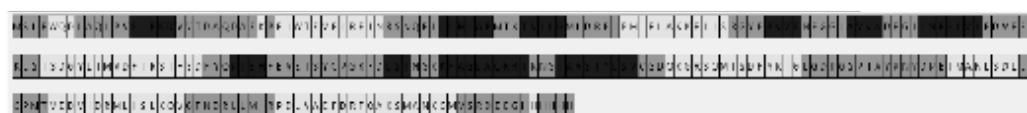


(C) 2P0L (Lipote-protein ligase A)

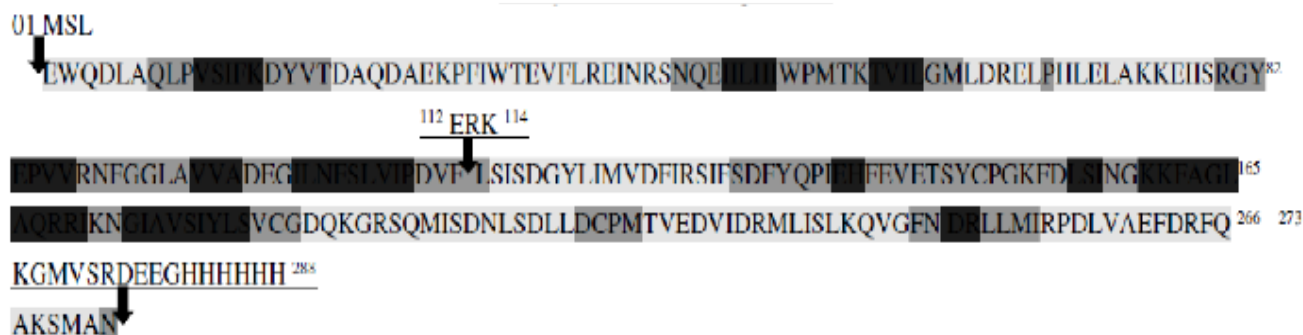
Mature protein

MSLEWQDLAQLPVSIKDYVTDAQDARKPFIWTEVFLREINRSNQEIIHWPMTKTIVILGMLDRELPHLEAKKEIISRGYBPVVRNFGGLAVVADE-  
GILKESLVIDVFTERKLSISDGYLIMVDFIRSIIFSDFYQPIIIEVETSYCPGKFDSLNGKKFAGLAQRRIK  
NGIAYSIYLSVCGDQKGRSQMISDIYKIGLGDTGSIPIAYPNVDPLIVANLSDLLDCPMIVEDVIDRMLISLQVGFNDRLLMIRPDLVAEFDREQAKSMANKGMVSRDEEGHHHHHH

Mature protein folded with JPred4



Crystal derived sequence



Crystal derived sequence file folded with JPred4

The image displays a sequence logo for a protein, with the title "Crystal derived sequence file folded with JPred4". The x-axis represents the sequence position, ranging from 1 to 250. The y-axis represents the information content in bits. The logo is divided into two main regions: a highly conserved N-terminal region (positions 1-100) and a less conserved C-terminal region (positions 101-250). The N-terminal region shows high conservation of specific residues, while the C-terminal region shows more variability.

**Manure oxycins**

Mature protein (folded with JIPred4)

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100 101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140 141 142 143 144 145 146 147 148 149 150 151 152 153 154 155 156 157 158 159 160 161 162 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 179 180 181 182 183 184 185 186 187 188 189 190 191 192 193 194 195 196 197 198 199 200 201 202 203 204 205 206 207 208 209 210 211 212 213 214 215 216 217 218 219 220 221 222 223 224 225 226 227 228 229 230 231 232 233 234 235 236 237 238 239 240 241 242 243 244 245 246 247 248 249 250 251 252 253 254 255 256 257 258 259 260 261 262 263 264 265 266 267 268 269 270 271 272 273 274 275 276 277 278 279 280 281 282 283 284 285 286 287 288 289 290 291 292 293 294 295 296 297 298 299 300 301 302 303 304 305 306 307 308 309 310 311 312 313 314 315 316 317 318 319 320 321 322 323 324 325 326 327 328 329 330 331 332 333 334 335 336 337 338 339 340 341 342 343 344 345 346 347 348 349 350 351 352 353 354 355 356 357 358 359 360 361 362 363 364 365 366 367 368 369 370 371 372 373 374 375 376 377 378 379 380 381 382 383 384 385 386 387 388 389 390 391 392 393 394 395 396 397 398 399 400 401 402 403 404 405 406 407 408 409 410 411 412 413 414 415 416 417 418 419 420 421 422 423 424 425 426 427 428 429 430 431 432 433 434 435 436 437 438 439 440 441 442 443 444 445 446 447 448 449 450 451 452 453 454 455 456 457 458 459 460 461 462 463 464 465 466 467 468 469 470 471 472 473 474 475 476 477 478 479 480 481 482 483 484 485 486 487 488 489 490 491 492 493 494 495 496 497 498 499 500 501 502 503 504 505 506 507 508 509 510 511 512 513 514 515 516 517 518 519 520 521 522 523 524 525 526 527 528 529 530 531 532 533 534 535 536 537 538 539 540 541 542 543 544 545 546 547 548 549 550 551 552 553 554 555 556 557 558 559 560 561 562 563 564 565 566 567 568 569 570 571 572 573 574 575 576 577 578 579 580 581 582 583 584 585 586 587 588 589 590 591 592 593 594 595 596 597 598 599 600 601 602 603 604 605 606 607 608 609 610 611 612 613 614 615 616 617 618 619 620 621 622 623 624 625 626 627 628 629 630 631 632 633 634 635 636 637 638 639 640 641 642 643 644 645 646 647 648 649 650 651 652 653 654 655 656 657 658 659 660 661 662 663 664 665 666 667 668 669 670 671 672 673 674 675 676 677 678 679 680 681 682 683 684 685 686 687 688 689 690 691 692 693 694 695 696 697 698 699 700 701 702 703 704 705 706 707 708 709 710 711 712 713 714 715 716 717 718 719 720 721 722 723 724 725 726 727 728 729 730 731 732 733 734 735 736 737 738 739 740 741 742 743 744 745 746 747 748 749 750 751 752 753 754 755 756 757 758 759 760 761 762 763 764 765 766 767 768 769 770 771 772 773 774 775 776 777 778 779 780 781 782 783 784 785 786 787 788 789 790 791 792 793 794 795 796 797 798 799 800 801 802 803 804 805 806 807 808 809 810 811 812 813 814 815 816 817 818 819 820 821 822 823 824 825 826 827 828 829 830 831 832 833 834 835 836 837 838 839 840 841 842 843 844 845 846 847 848 849 850 851 852 853 854 855 856 857 858 859 860 861 862 863 864 865 866 867 868 869 870 871 872 873 874 875 876 877 878 879 880 881 882 883 884 885 886 887 888 889 890 891 892 893 894 895 896 897 898 899 900 901 902 903 904 905 906 907 908 909 910 911 912 913 914 915 916 917 918 919 920 921 922 923 924 925 926 927 928 929 930 931 932 933 934 935 936 937 938 939 940 941 942 943 944 945 946 947 948 949 950 951 952 953 954 955 956 957 958 959 960 961 962 963 964 965 966 967 968 969 970 971 972 973 974 975 976 977 978 979 980 981 982 983 984 985 986 987 988 989 990 991 992 993 994 995 996 997 998 999 1000 1001 1002 1003 1004 1005 1006 1007 1008 1009 1010 1011 1012 1013 1014 1015 1016 1017 1018 1019 1020 1021 1022 1023 1024 1025 1026 1027 1028 1029 1030 1031 1032 1033 1034 1035 1036 1037 1038 1039 104

Crystal derived sequence

01 MVHHHHHHSSGHEHTG 15  
↓  
QSAAATMPLDSIDVSIPELFYNDISVGEYFKRLRKDDPFIYCADSAFGPQVSTKYNDIMHVDTNHDIISDAGYGG 94  
98 QKGGDGG 105  
↓  
IIDDGILDLPNFIAMDRPRHDEQRKAVSPIVAPANLAALLEGITIRERVSKTLDGLPVGEEIIVWDRVSIETTQMLATLFI 178  
DIPFEERRKLTRWSDVTTAAPGGGGVVEISWDQRKTELLECAAYFQVLWNERVNKDPCNDLISMLAHSPATRNMTPE 253  
EYLGNIJLLIVGGNDTTRNSMTGGVIALHKNPDQFAKIKANPAIVETMVPPIIRWQTPLAHIDGDTADSGGG 329  
RKGDVWYYSGNRDDEVIDRPEEPIIDRPRPRQHLSPGIGIHIRCVGNRLAEMQLRILWEEILTRFSKQMAEP 406  
VRSNFVRGMA 426

Crystal derived sequence folded with JPred4

[illegible]

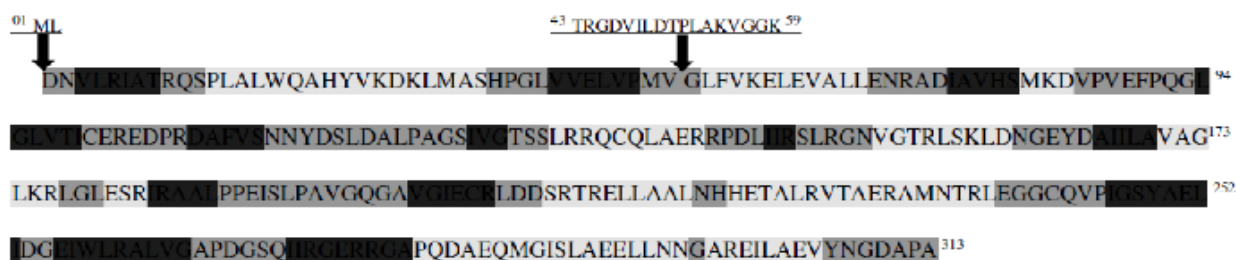
## Mature protein

Mixture protein folded with IPred4

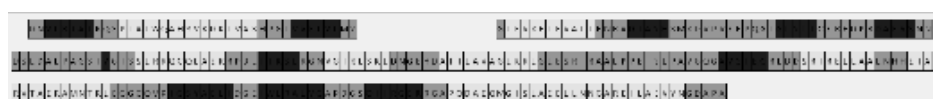
[illegible]

Crystal derived sequence

(Figure 1 continues)



Crystal derived sequence folded in JPred4

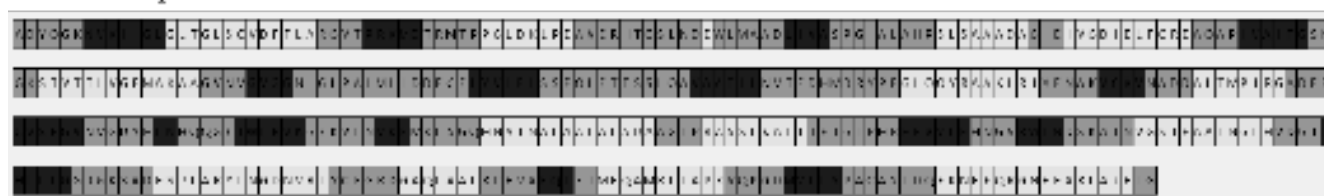


## (F) 1UAG (UDP-N-acetylmuramoyl-L-alanine:D-glutamate ligase)

Mature protein

ADYQGKNVVTIGLGLTGLSCVDFFLARGVTPRVMDTRMTTPGLDKLPEAVERHTGSLNDEWTLMAADLIVASPGIALAHPSLSAAADAGILVGDIELFCRELAQAPIVAHFGSNGKSTVTTLVGEMAKAAGVNVGVGGNIGLPALMLLDDECFYVLELSSFQLETTSSLSQAVAATHLVNTEITMDRYPFGLQQYRAAKLRIVENAKVCVVAADDALTMPIRGADIERCVSFQVVMGQDYHLNHOOGETWLRVKGEKVLNVKEMKLSGQHNYINATAALALADAAGLPRASSLKALTTTGLPIIRFVLEHNGVRWINDSKATNVGSTEAALNGLHVDGTLHLHLOGDGKSADESPRARYLNGDNVRLYCTGRDGAQLAALRPEVALQTETMEQAMRLLAPRVQPGDMVTLSPACASLDQFKNFEQRGNEFARLAKELG

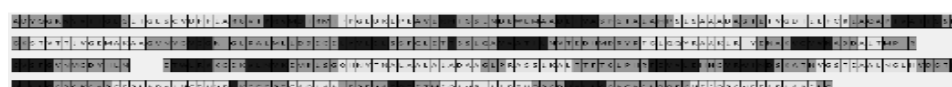
Mature protein folded with JPred4



Crystal derived sequence

ADYQGKNVVTIGLGLTGLSCVDFFLARGVTPRVMDTRMTTPGLDKLPEAVERHTGSLNDEWTLMAADLIVASPGIALAHPSLSAAADAGILVGDIELFCRELAQAPIVAHFGSNGKSTVTTLVGEMAKAAGVNVGVGGNIGLPALMLLDDECFYVLELSSFQLETTSSLSQAVAATHLVNTEITMDRYPFGLQQYRAAKLRIVENAKVCVVAADDALTMPIRGADIERCVSFQVVMGQDYHLNHOOGETWLRVKGEKVLNVKEMKLSGQHNYINATAALALADAAGLPRASSLKALTTTGLPIIRFVLEHNGVRWINDSKATNVGSTEAALNGLHVDGTLHLHLOGDGKSADESPRARYLNGDNVRLYCTGRDGAQLAALRPEVALQTETMEQAMRLLAPRVQPGDMVTLSPACASLDQFKNFEQRGNEFARLAKELG<sup>437</sup>

Crystal derived sequence folded with JPred4





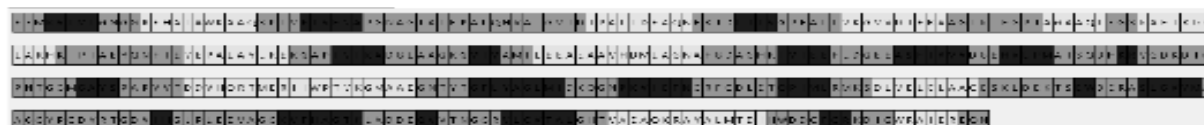
(Figure 1 continues)

## (G) IGSO (Glycinamide ribonucleotide synthase)

Mature protein

EFMKVLVIGNGGREHALAWKAAQSPVETVTVAPGNAGTALBPALQNVAGVTDIPALLDFAQNEKID  
 LITVGIPLAPLVKGVVDITFRAAGIKIPGPTAGAAQHLGSKAFIKDITLARHKIPTALYQNFITVEPALAYL  
 REKGAPIVKA **DGLAAG** KGVIVAMTLEEAEEAAVHDMLAGNAPGDAGHRVIEEFLDGEESFIVMVDG  
 CHVLPMAISQIRIKLVGDKDITGPNITGGMGAAYSPAPVVTDDVHQRTMERIIWPTVKGMAAEGNTYIGI  
 LYAGLMIDKQGNPKVIEFNCRPGDLETQPIMLRMKSDLVELCLAACESKLDEKTSWDERASLGVM  
 AAGGYPGDYRTGDIHGLPIEEVAGG **DDF** AGTKLA **QVVTNGG**  
 ALMTDIHWDDCFCKDIGWRAIER **EON**

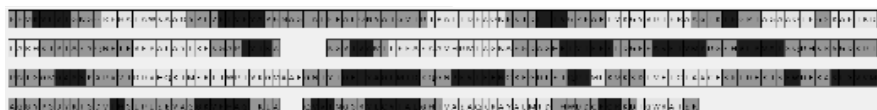
Mature protein folded with JPRED4



Crystal derived sequence

EFMKVLVIGNGGREHALAWKAAQSPVETVTVAPGNAGTALBPALQNVAGVTDIPALLDFAQNEKID **DDF** GFEA 76  
 118 **DGLAAG** 154  
 PLVKGVDITFRAAGIKIPGPTAGAAQLEGSKAFIKDITLARHKIPTALYQNFITVEPALAYLREKGAPIVKA **KG** 158  
 MTLEEAEEAAVHDMLAGNAPGDAGHRVIEEFLDGEESFIVMVDGCHVLPMAISQIRIKLVGDKDITGPNITGGM 232  
 SPAPVVTDDVHQRTMERIIWPTVKGMAAEGNTYIGI LYAGLMIDKQGNPKVIEFNCRPGDLETQPIMLRMKSD 308  
 374 **DDF** 377  
 LVELCLAACESKLDEKTSWDERASLGVM AAGGYPGDYRTGDIHGLPIEEVAGG **DDF** AGTKLA **QVVTNGG**  
 428 **EON** 431  
 VLVN **ALGH**IVAEAQKRAYALMTDIHWDDCFCKDIGWRAIER

Crystal derived sequence file folded with JPRED4

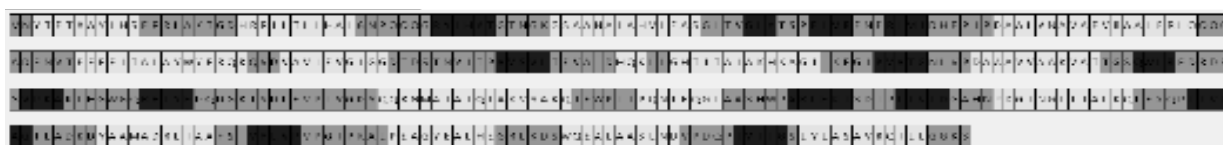


## (H) IFGS (Folypolyglutamate synthetases)

Mature protein

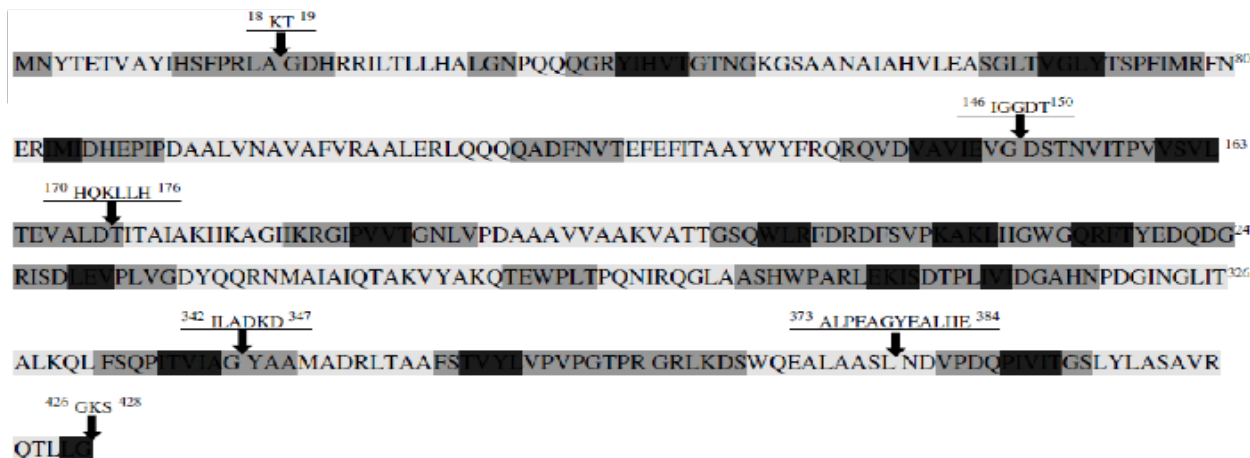
VNYTETVAYTHSEPRLA **K** TGDITRRILTLITIALGNPQQQGRVTHVTGTNGKGSAAANATAITVLEASGLTV  
 GLYTSPFLMRI NLRIMIDHLPIDAAALYNAAVAVRAALLERLQQQXADINVTLEITIALAYWYLRQRQV  
 DVAVIRVGH **GGD** I DSEINVTVPVSVITFEVALD **RQKLEGH** ITTAAKHKAGHKRGIPVVTGNIAPDAAAV  
 VAAKVATTGSQWLRFD RDFSVPKAKLHGWWGQRFTYEDQDGRISDLEVPLVGDYQQRNMAIAIQTA  
 VYAKQITWPLIPONIROGLAASIWPARLEKISDTPLIVHDGAINPDGINCHITIAKQLDSQPIVIAG **LA**  
 DKDYAAMADRLTAASFVYLVVPVPGTPRALPEAGYEAL **HE** GRLKDSWQEAALAASLNDVPDQPIVITGS  
 LYLASAVRQILL **GGSS**

Mature protein folded with JPRED4

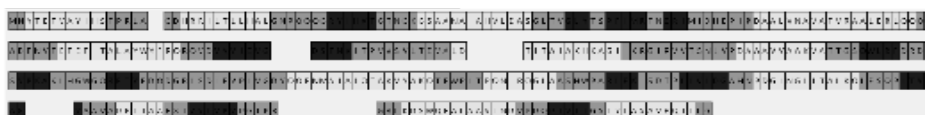


(Figure 1 continues)

Crystal derived sequence 3.1e



Crystal derived sequence 3.1e folded with JPred4



## (f) 3RC3 (Human Mitochondrial Helicase Suv)

Mature protein

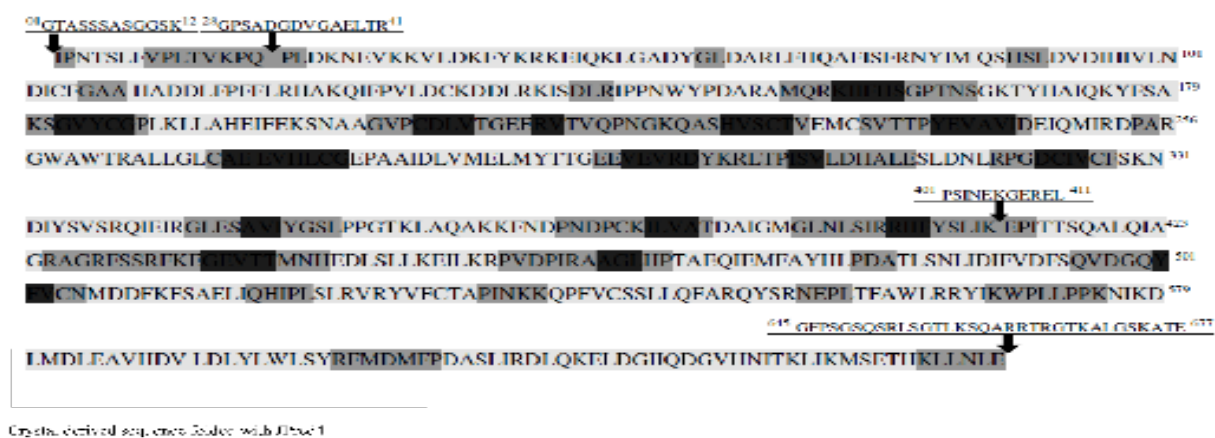
GIASSASAGGSKIPNLSLEVPILVVKIPQGI<sup>1</sup>SAIDHNGAELIR<sup>2</sup>PIIDKNEVKKVLDKFKYKRKFIQKIGADYGLDARLPHQAFISERNYIMQSHSLDVEDIHVLNDICIGAAHADDLPFETLRHAKQIFPVLDCKDDLRKISIDRIHPTNWYPIDARAMQKRIHUSGPTNSGKIYTHAIQKYPSAKSGIVYCGPIKILLATHIHIRSNAAAGVPCIDLVLTGEERTVTQPNGKQASHVSCIVEMCSVTTPYEVAVIDEIQMIRDPARGWAWTRALLGLCAEEVHLGGTPAAIDLVMTIMYTTGTFVEVRDYKRIITPISVLDTIALESIDNLRPGDCTVCFSENDEYSSVRQIEIRGLASAVIYGSGLPPTGKTAKAKKFNDDPNDCIKILVATDAIGMGLNLSIRRIIFYSLIKPSINEKGEREL<sup>3</sup>EPITTSQALQIAGCAGAGRESSRFKFGVTTIMHIDLSLKEIKRPVDPIRAAGLIHPTAQIQRMEAYTHLPDAHISNLIDIFVDFSQVDGQYFVCNMDDFKFSAEIQLHIPLSLRVRYVECTAPINKKQPFVCSLLQFARQYSRNEPLIFAWLRRYIKWPELPPKNIKIDIMDLAVIHDYLDLYLWLSYKEMDMFEDASLRIDIQKELIDGHIQDG VHNITKLIKMSETHKLLNLLGI<sup>4</sup>PSGISOSSRLSGTLKSOARRTRGTGKALGSKATE<sup>5</sup>

Mature protein folded with JPred4



Crystal derived sequence

(Figure 1 continues)



**Figure 1** Panorama of secondary structures from mature protein sequence (FASTA) and crystal structure sequence folded with JPred4: The different secondary domains are coloured according to gray scale. The crystal derived sequence is also colour coded according to secondary structures in grey scale and the arrows indicate the missing region. The missing oligopeptide region in crystal derived sequence is highlighted in the FASTA sequence.

**Table 1** Missing oligopeptide in the crystal structure derived sequence

Protein	PDB Id	Missing residues in crystal derived sequence		
		N terminal	C terminal	internal strings
Saicar synthase	1A48	1	0	7
Saicar synthase	2POL	3	16	3
Lipoate-protein ligase A	3RWL	15	0	7
P450 pyr hydroxylase	2YPN	2	0	17
Hydroxymethylbilane synthase	2CNQ	1	0	3
UDP-n-acetylmuramoyl- Lalanine: D-glutamate ligase	1UAG	0	0	5, 4
Glycinamide ribonucleotide synthetase	1GSO	0	0	6, 3
Folypolyglutamate synthetase	1FGS	0	0	32, 5, 7, 6, 12
Mitochondrial helicase suv3	3RC3	12	0	14, 11, 33



The secondary structures predicted for internal missing oligopeptide and their flanking tripeptides from both FASTA and PyMol (crystal) formats are shown in Table 3. Surprisingly, 10 out of 16 internal missing oligopeptides form the disordered domains (DD). Among the remaining 6, two disordered domains adjoin terminal residue from  $\beta$  sheet, one adjoins  $\alpha$  helix and 3 are from putative helix. In the tripeptides flanking the internal missing strings, we find that,

at N-terminal, 10 out of 16 forms IDD, 3 form beta sheets 2 are from  $\alpha$  helix and 1 form a junction between beta sheet and random coil. In the C terminal tripeptide, 7 are from disordered domain, 3 form a junction between disorder domain and  $\alpha$  helix, 2  $\beta$  sheet- DD junctions and 2 each from  $\beta$  sheet and  $\alpha$ - helix. Thus, clearly, all missing strings are part of original disordered domains

**Table 2** Distribution of missing residues in crystal structure

Protein	PDB Id	Missing oligopeptide	Proline residues	glycine residues	charged residue	Polar uncharged	hydro-phobic	Total amino acid
Saicar synthase	1A48	KAEQGEH	0	1	4	1	2	7
Saicar synthase	2CNQ	EQG	0	1	1	1	1	3
Lipoate-protein ligase A	2POL	ERK	0	0	3	0	0	3
P450 pyr hydroxylase	3RWL	QKGGDGG	0	4	2	1	4	7
Hydroxymethylbilane synthase	2YPN	TRGDVILDTPLAKVGGK	1	3	5	2	10	17
UDP-n-acetylmuramoyl-L-alanine-D-glutamate ligase	IUAG	GADER	0	1	3	0	2	5
"	"	HQQG	0	1	1	2	1	4
Glycinamide ribonucleotide synthetase	1GSO	DGLAAG	0	2	1	0	5	6
"	"	DDE	0	0	3	0	0	3
Folypolyglutamate synthetase	IFGS	KT	0	0	1	1	0	2
"	"	IGGDT	0	2	1	1	3	5
"	"	HQKLLGH	0	1	3	1	3	7
"	"	ILADKD	0	0	3	0	3	7
"	"	ALPEAGYEALHE	1	1	4	0	7	12
Mitochondrial Helicase sub 3	3RC3	GPSADGDVGAELTR	0	3	4	2	8	14
"	"	PSINEKGEREL	1	1	5	2	4	11

Note: only one aromatic residue (tyrosine) was seen in the internal missing oligopeptide string in IFGS.

**Table 3** Predicted secondary structure of the internal missing oligopeptide and the flanking residues

Protein name	Protein Id	Missing oligopeptide	Secondary structure of residues after folding FASTA sequence in JPred4		
			missing	Tripeptide flanking the missing oligopeptide region	
				N terminal	C terminal
SAICAR synthase	IA48	KAEQGEH	random coil	random coil	random coil
SAICAR synthase	2CNQ	EQG	random coil	random coil	random coil
Lipoate-protein ligase A	2POL	ERK	random coil	random coil	random coil and $\alpha$ helix
P450 pyr hydroxylase	3RWL	QKGGDGG	random coil	random coil	random coil
Hydrox-ymethylbilane synthase	2YPN	TRG DVILDTPLAKVGGK	$\beta$ sheet and random coil	$\beta$ sheet	random coil and $\alpha$ helix
UDP-n-acetylmuramoyl L-alanine D-glutamate ligase	IUAG	GAD ER	$\beta$ sheet and random coil	$\alpha$ helix	$\beta$ sheet
"	"	HQQG	random coil	$\beta$ sheet	$\beta$ sheet
Glycinamide ribonucleotide synthetase	IGSO	DGLAAG	random coil	$\beta$ sheet and random c oil	random coil
"	"	DDE	random coil	random coil	random coil and $\beta$ sheet
Folypolyglutamate synthetase	IFGS	KT	random coil	random coil	random coil and $\alpha$ helix
"	"	IGGDT	$\alpha$ helix and random coil	$\alpha$ helix	random coil
"	"	HQKLLGH	$\alpha$ helix and random coil	random coil	$\alpha$ helix
"	"	ILADKD	random coil	$\beta$ sheet	$\alpha$ helix
"	"	ALPEAGYEALHE	$\alpha$ helix and random coil	random coil	random coil
Mitochondrial Helicase sub 3	3RC3	GPSADGDVGAELTR	random coil	random coil	random coil
"	"	PSINEKGEREL	$\beta$ sheet and random coil	random coil	$\beta$ sheet and random coil

Comparing panoramas of secondary structures derived from the crystal structure to those computed by folding sequences from both, mature proteins and crystals with JPred4, we find (Table 4) that for each type of secondary structure crystals give an underestimate of the number of disordered domains as well as the number of residues therein. Indeed, a combined analysis of 9 proteins reveal the ratio (number of secondary structure domains: number of amino

acid residues) is comparable for  $\alpha$  helices and  $\beta$  sheets, but substantially reduced for disordered domains in crystals than *in silico* folded mature protein. Similar results were obtained by folding with PSSpred (not shown).

We find that only 2 crystals reveal histidine-rich oligopeptides at N or C terminal and none in the internal missing oligopeptide strings (data not shown).

**Table 4** Secondary structure from the mature protein sequence and crystal structure derived sequence folded using Jpred4 and the original crystal sequence

Protein name	PDB ID	Source	Alpha Helix [no. of motifs (no. of amino acids)]			Beta sheets [no. of motifs(no. of amino acids)]			Random coils [no. of motifs(no. of amino acids)]		
			mature protein		crystal structure	mature protein		crystal structure	mature protein		crystal structure
			JPred4	derived		Jpred	derived		JPred4	derived	JPred4
Saicar synthase	1A48	<i>Saccharomyces cerevisiae</i> ATCC 204508	6 (82)	7 (116)	7(85)	10 (56)	15 (106)	8 (53)	17(168)	20 (84)	16(160)
Saicar synthase	2CNQ	<i>Saccharomyces cerevisiae</i> ATCC 204508	6 (82)	6 (127)	5(82)	10 (56)	15 (69)	10 (57)	17(168)	17 (106)	16(163)
Lipoate protein ligase A	2POL	<i>Sphingopyxis macrogoltabida</i>	10 (105)	9 (118)	8(93)	10 (57)	11 (55)	10 (59)	21(126)	20 (93)	19(114)
P450 pyr Hydroxylase	3RWL	<i>Sphingopyxis macrogoltabida</i>	14 (76)	14 (234)	14(177)	6 (36)	12 (40)	6 (34)	19(214)	24 (130)	20 (193)
Hydroxymethylbilane synthase	2YPN	<i>Escherichia coli</i> K12	8 (112)	11 (112)	8 (113)	11 (69)	13 (76)	11(62)	20(132)	21 (106)	20(119)
UDP-n-acetylmuramoyl L-alanine D-glutamate ligase	IUAG	<i>Escherichia coli</i> K12	15 (161)	20 (161)	20(152)	17 (83)	20 (89)	20(88)	32(193)	38 (178)	33(188)
Glycinamide ribonucleotide synthetase	IGSO	<i>Escherichia coli</i> K12	11 (127)	16 (128)	12(130)	20 (97)	16 (99)	12(99)	32(207)	33 (192)	32(190)
Folypolyglutamate synthetase	IFGS	<i>Lactobacillus casei</i>	16 (192)	15 (172)	13(166)	14 (67)	16 (62)	13(70)	31(169)	29 (159)	27(157)
Mitochondrial helicase suv3	3RC3	<i>Homo sapiens</i>	31 (441)	26 (444)	26(366)	13 (60)	16 (69)	14(65)	43(176)	42 (164)	37(246)

Table 2 lists the relative distribution of Proline, Glycine, charged and hydrophobic residues in the internal missing strings. Thus, there is a high concentration of flexible (Glycine, 20/107) and charged residues (43/107) in these strings, while rigid Proline is of rare occurrence (3/107). Similarly, there is only 1 aromatic residue in the missing strings (not shown).

## Discussion

According to Djinovic-Carrugo & Carrugo,<sup>1</sup> most crystallographic data reveal incidence of internal missing strings of oligopeptides. Here we describe in detail 9 such strings and analyses of their position in the overall panorama of secondary structure domains of a polypeptide sequence. To study this aspect we have adopted the strategy of folding *in silico* sequences for the same protein representing the post-translationally processed polypeptide and that derived from the crystal. The issue here is that when the crystal structure is obtained at low resolution, a number of residues fail to be detected due to low electron density. Therefore, by comparing two amino acid sequences of the same protein, we find the missing residues missing in crystal-derived sequence.

Comparison of the panorama of secondary structure domains revealed that after folding the sequences *in silico*, allowed us to detect secondary structure domains to which the missing residue belong and we conclude that most are disordered domains. This is further supported by the fact that these are rich in flexible amino acid Glycine and poor in rigid Proline. We find that out of the 16 cases, the proportion of hydrophobic residue is less than 0.5 in 10 cases and in remaining, it is less than 0.6. These disordered oligopeptides contain high concentration of charged residues and nearly 20% glycines. We conclude that the apparent loss or delectability in crystals of large internal oligopeptide strings involve highly disordered domains which probably accounts for the difficulty crystallographers face in designating a signature domain to the missing internal string. In fact, since these strings are not actually absent in polypeptides, the inability to detect leads to an incomplete crystal structure. Clearly, in most cases the problem can be solved by comparing the *in silico* folded amino acid sequences of mature proteins to those derived from crystals.

Finally, one must consider the structural and functional relevance of the apparently missing segments. To that effect we are now assessing the propensity of various Triads involved in defining

functionally important sites for enzyme-substrate interactions as well as other protein: protein binding. Another possible approach is to examine the missing oligopeptides alone and with flanking regions in Ramachandran plots. The question, therefore, remains as to how one should solve the crystal structure beyond the offerings of crystallography. In any case, it is unlikely that proteins exist in crystalline form in vivo and probably exhibit a metastable state with variable mobility of flexible regions depending on the intracellular environment.

Indeed, polymeric structures, namely, micelles, membranes and globular proteins exhibit a hydrophobic core and hydrophilic exterior that are differentially sensitive to perturbations by osmotic pressure, ionic strength and temperature and exhibit differential movements such that the kinetic energy between the two domains is conserved.<sup>9</sup>

## Acknowledgments

We are grateful to Prof. C. Ramakrishnan, Prof. T. Ramasarma, Prof. N.V. Joshi and Prof. V. Sitaramam for critical comments and suggestions. We also thank Dr. Vijayanti Tamhane and Prof. Ameeta Ravi Kumar, Head, IBB, for encouragement and support.

## Conflict of interest

The author declares that there is no conflict of interest regarding the publication of this article.

## References

1. Djinovic-Carugo K, Carugo O. Missing strings of residues in protein crystal structures. *Intrinsically Disord Proteins*. 2015;3(1):e1095697.
2. Modak SP, Milner Kumar M, Bargaje. *Molecular Phylogenetic Trees: Topology of multiparametric poly-genic/phenic Tree exhibit Higher Taxonomic Fidelity than Uniparametric Trees for Mono-Genic/Phenic traits in Evolutionary Biology: Mechanisms and trends*. Springer verlag; 2012. p. 79–102.
3. www.rcsb.org
4. Levdikov VM, Barynin VV, Grebenko AI, et al. The structure of SAICAR synthase: an enzyme in the de novo pathway of purine nucleotide biosynthesis. *Structure*. 1983;6(3):363–376.
5. Pham SQ, Pompidor G, Liu J, et al. Evolving P450<sub>pyr</sub> hydroxylase for highly enantioselective hydroxylation at non-activated carbon atom. *Chemical Communications*. 2012;48(38):4618–4620.
6. Nieh YP, Raftery J, Weisgerber S, Habash J, et al. Accurate and highly complete synchrotron protein crystal Laue diffraction data using the ESRF CCD and the Daresbury Laue software. *Journal of Synchrotron Radiation*. 1999;6:995–1006.
7. Bertrand JA, Auger G, Fanchon E, et al. Crystal structure of UDP-N-acetylmuramoyl-L-alanine: D-glutamate ligase from *Escherichia coli*. *EMBO J*. 1997;16(12):3416–3425.
8. Wang Weiru T, JosephKappock, JoAnne Stubbe, et al. “X-ray crystal structure of glycinamideribonucleotide synthetase from *Escherichia coli*.” *Biochemistry*. 1998;37(45):15647–15662.
9. Madhavarao CN, Sauna ZE, Srivastava A, et al. Osmotic perturbations induce differential movements in the core and periphery of proteins, membranes and micelles. *Biophys Chem* 2011;90(3):233–248.
10. Sun X, Bognar AL, Baker EN, et al. Structural homologies with ATP-and folatebinding enzymes in the crystal structure of folylpolyglutamatesynthetase. *Proc Natl Acad Sci U S A*. 1998;95(12):6647–6652.
11. Jedrzejczak R, Wang J, Dauter M, et al. Human Suv3 protein reveals unique features among SF2 helicases. *Acta Crystallographica Section D: Biological Crystallography*. 2011;67(11):988–996.
12. <http://zhanglab.cmb.med.umich.edu/PSSpred/>
13. Drozdetskiy A, Cole C, Procter J, et al. JPred4: a protein secondary structure prediction server. *Nucleic Acids Res*. 2015;43(W1):W389–94.
14. Dunker AK, Babu MM, Barbar E, et al. What’s in a name? Why these proteins are intrinsically disordered: Why these proteins are intrinsically disordered. *Intrinsically Disordered Proteins*. 2013;(1):e24157.
15. <http://www.compbio.dundee.ac.uk/jpred/>