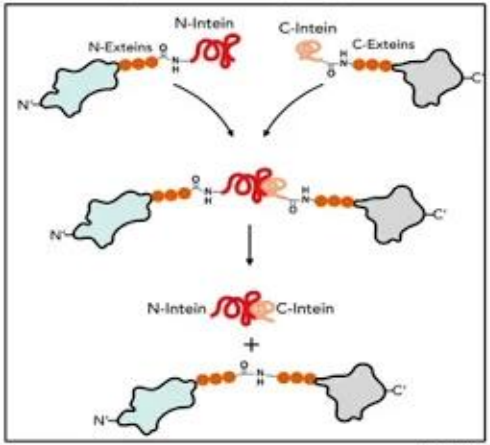
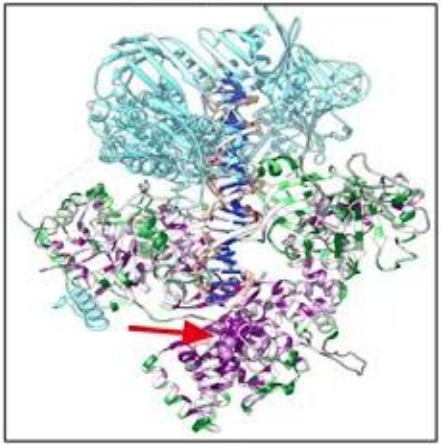
Split Inteins

Split Intein Self splicing



Barzel 2011

DnaE (Extein) Protein Structure



1
00:00:04,550 --> 00:00:02,230
hello my name is daniel phillips i'm a

2
00:00:06,389 --> 00:00:04,560
graduate student in the go garden lab

3
00:00:07,909 --> 00:00:06,399
part of the mcb department in the

4
00:00:10,110 --> 00:00:07,919
university of connecticut

5
00:00:14,390 --> 00:00:10,120
and this is my talk on split intense and

6
00:00:18,070 --> 00:00:15,829
so first of all i would like to stress

7
00:00:19,910 --> 00:00:18,080
the important role cyanobacteria play

8
00:00:21,510 --> 00:00:19,920
in the evolution of uh earth and the

9
00:00:24,070 --> 00:00:21,520
evolution of life on earth

10
00:00:24,790 --> 00:00:24,080
were we to go back billions of years ago

11
00:00:26,390 --> 00:00:24,800
earth

12
00:00:28,070 --> 00:00:26,400
had a very different atmosphere it had a

13
00:00:30,150 --> 00:00:28,080

reducing atmosphere there was very

14
00:00:31,910 --> 00:00:30,160
little o2 uh

15
00:00:33,270 --> 00:00:31,920
that changes when we start to see

16
00:00:36,549 --> 00:00:33,280
cyanobacteria

17
00:00:38,069 --> 00:00:36,559
uh around uh two and a half two billion

18
00:00:40,229 --> 00:00:38,079
years ago where we start to see

19
00:00:41,350 --> 00:00:40,239
a gradual and then rapid rise in the

20
00:00:43,590 --> 00:00:41,360
concentration

21
00:00:44,549 --> 00:00:43,600
of o2 in earth's atmosphere until the

22
00:00:47,270 --> 00:00:44,559
level uh

23
00:00:47,670 --> 00:00:47,280
that it's at today around 20 percent and

24
00:00:49,750 --> 00:00:47,680
then

25
00:00:51,270 --> 00:00:49,760
even today around 50 of the earth's

26
00:00:53,229 --> 00:00:51,280
atmospheric oxygen

27
00:00:55,510 --> 00:00:53,239
uh can be traced back to uh

28
00:00:57,029 --> 00:00:55,520
cyanobacteria uh in the ocean

29
00:01:00,630 --> 00:00:57,039
um you know making them the true lungs

30
00:01:04,229 --> 00:01:02,389
and so in trying to learn more about the

31
00:01:05,270 --> 00:01:04,239
evolution of cyanobacteria i'm

32
00:01:08,070 --> 00:01:05,280
interested

33
00:01:09,350 --> 00:01:08,080
in the distribution and function of

34
00:01:13,270 --> 00:01:09,360
intense

35
00:01:14,230 --> 00:01:13,280
within this phylum intenes are genetic

36
00:01:16,870 --> 00:01:14,240
elements

37
00:01:19,109 --> 00:01:16,880
that reside inside important

38
00:01:22,630 --> 00:01:19,119

housekeeping genes

39
00:01:27,590 --> 00:01:25,990
in the genome so we would see in the top

40
00:01:30,069 --> 00:01:27,600
of this figure here

41
00:01:32,550 --> 00:01:30,079
we would see in red the if this is the

42
00:01:34,469 --> 00:01:32,560
dna sequence the genomic dna

43
00:01:36,550 --> 00:01:34,479
we will see the in teen coding sequence

44
00:01:37,270 --> 00:01:36,560
sandwiched between the extine the extene

45
00:01:40,390 --> 00:01:37,280
would be

46
00:01:42,710 --> 00:01:40,400
the host protein

47
00:01:43,510 --> 00:01:42,720
or the dna which would code for the host

48
00:01:46,149 --> 00:01:43,520
protein

49
00:01:46,550 --> 00:01:46,159
uh the cnn terminal of that that would

50
00:01:50,630 --> 00:01:46,560
be

51
00:01:53,990 --> 00:01:50,640
an rna

52
00:01:57,830 --> 00:01:54,000
on our on rna molecule uh combining

53
00:01:59,270 --> 00:01:57,840
of the nxt the intent and the cex team

54
00:02:00,630 --> 00:01:59,280
and then that will be translated

55
00:02:02,230 --> 00:02:00,640
together so now we have this long

56
00:02:04,230 --> 00:02:02,240
polypeptide chain

57
00:02:06,310 --> 00:02:04,240
uh containing the extine the host

58
00:02:09,669 --> 00:02:06,320
protein and the intent

59
00:02:12,070 --> 00:02:09,679
the intent will uh splice itself out

60
00:02:13,510 --> 00:02:12,080
uh spontaneously out of this uh

61
00:02:15,350 --> 00:02:13,520
polypeptide chain

62
00:02:16,790 --> 00:02:15,360
and that will leave the mature host

63
00:02:23,030 --> 00:02:16,800

protein and the now

64
00:02:26,470 --> 00:02:23,040
excised in teen

65
00:02:28,790 --> 00:02:26,480
so they're distributed fairly widely

66
00:02:32,470 --> 00:02:28,800
among uh the three domains of life

67
00:02:37,589 --> 00:02:35,270
we see them uh pretty widespread and uh

68
00:02:40,150 --> 00:02:37,599
in in bacteria and archaea

69
00:02:41,350 --> 00:02:40,160
and eukaryota they are limited to fungi

70
00:02:44,630 --> 00:02:41,360
in protists

71
00:02:48,309 --> 00:02:44,640
limited to single celled

72
00:02:48,949 --> 00:02:48,319
uh organisms and pretty much entirely

73
00:02:55,430 --> 00:02:48,959
they do

74
00:03:02,309 --> 00:02:59,270
so chloroplasts but they are absent from

75
00:03:05,910 --> 00:03:02,319
the uh nuclear genome

76
00:03:07,670 --> 00:03:05,920
of uh of all multicellular uh domains of

77
00:03:09,750 --> 00:03:07,680
life

78
00:03:11,430 --> 00:03:09,760
so there are three main types of in

79
00:03:12,309 --> 00:03:11,440
teams first we have the full length

80
00:03:15,270 --> 00:03:12,319
enteine

81
00:03:18,070 --> 00:03:15,280
uh this will contain uh the in this will

82
00:03:19,589 --> 00:03:18,080
contain a homing endonuclease domain uh

83
00:03:21,910 --> 00:03:19,599
so if you can imagine this will be

84
00:03:25,030 --> 00:03:21,920
translated uh transcribed the in team

85
00:03:27,030 --> 00:03:25,040
will self-splice itself the nsr as csr

86
00:03:28,390 --> 00:03:27,040
and terminal splicing region c terminal

87
00:03:30,229 --> 00:03:28,400
splicing you can splice itself

88
00:03:31,990 --> 00:03:30,239

out and then it can go back into the

89
00:03:33,990 --> 00:03:32,000
genome and look for a

90
00:03:35,750 --> 00:03:34,000
non-in teen containing allele if it

91
00:03:38,789 --> 00:03:35,760
finds one it will cut that

92
00:03:39,190 --> 00:03:38,799
um dna strand and then the host cell

93
00:03:41,910 --> 00:03:39,200
will

94
00:03:43,350 --> 00:03:41,920
repair it using the dna can using the

95
00:03:45,910 --> 00:03:43,360
incontaining allele

96
00:03:48,070 --> 00:03:45,920
as a template and this allows the intent

97
00:03:49,110 --> 00:03:48,080
to maintain itself in a genome making it

98
00:03:52,229 --> 00:03:49,120
much more difficult

99
00:03:54,229 --> 00:03:52,239
to be lost

100
00:03:56,070 --> 00:03:54,239
uh the mini antene of course uh will

101
00:03:59,270 --> 00:03:56,080
have uh no will have lost its

102
00:04:02,309 --> 00:03:59,280
endonuclease uh containing domain

103
00:04:03,830 --> 00:04:02,319
uh if it ever had one uh and so once

104
00:04:05,190 --> 00:04:03,840
lost from the genome it's unable to

105
00:04:08,309 --> 00:04:05,200
reinsert itself

106
00:04:10,149 --> 00:04:08,319
uh back in and then the split in team on

107
00:04:12,309 --> 00:04:10,159
what i am mostly interested in

108
00:04:13,429 --> 00:04:12,319
uh will have the n terminal and the c

109
00:04:17,030 --> 00:04:13,439
terminal uh

110
00:04:19,030 --> 00:04:17,040
located in um different loci

111
00:04:21,430 --> 00:04:19,040
of the genome and this could be tens of

112
00:04:23,510 --> 00:04:21,440
thousands of base pairs apart

113
00:04:24,870 --> 00:04:23,520

uh so you can imagine how baroque this

114
00:04:28,469 --> 00:04:24,880
process is

115
00:04:29,510 --> 00:04:28,479
um where uh we have these basically

116
00:04:32,790 --> 00:04:29,520
non-functioning

117
00:04:34,310 --> 00:04:32,800
to the host cell uh pep uh uh

118
00:04:36,870 --> 00:04:34,320
genes that are gonna be transcribed

119
00:04:38,310 --> 00:04:36,880
separately uh translated separately

120
00:04:40,790 --> 00:04:38,320
and then we have these non-functioning

121
00:04:42,710 --> 00:04:40,800
peptide chains um that have to go

122
00:04:43,670 --> 00:04:42,720
through the cytoplasm until the intense

123
00:04:46,710 --> 00:04:43,680
can link up

124
00:04:49,749 --> 00:04:46,720
combine uh excise themselves

125
00:04:50,390 --> 00:04:49,759
um connect the two host x teams host

126
00:04:52,150 --> 00:04:50,400
proteins

127
00:05:00,230 --> 00:04:52,160
uh giving the native mature functional

128
00:05:03,990 --> 00:05:01,990
so i'd like to talk more about uh the

129
00:05:08,310 --> 00:05:04,000
the split in teen particularly the dna

130
00:05:10,469 --> 00:05:08,320
e uh split in teen in cyanobacteria

131
00:05:12,150 --> 00:05:10,479
and so on the left we see an image that

132
00:05:12,950 --> 00:05:12,160
demonstrates a little bit more in depth

133
00:05:15,830 --> 00:05:12,960
of how

134
00:05:16,310 --> 00:05:15,840
uh the n terminal and the c terminal of

135
00:05:23,189 --> 00:05:16,320
the

136
00:05:25,029 --> 00:05:23,199
so as i said before this is a little bit

137
00:05:27,270 --> 00:05:25,039
of a complex process

138
00:05:28,870 --> 00:05:27,280

process and so it makes one think what

139
00:05:32,070 --> 00:05:28,880
benefit this would have

140
00:05:32,629 --> 00:05:32,080
uh to an organism well that's some quite

141
00:05:39,909 --> 00:05:32,639
uh

142
00:05:41,110 --> 00:05:39,919
these in teens have a role or not there

143
00:05:44,310 --> 00:05:41,120
are some researchers

144
00:05:47,110 --> 00:05:44,320
who think they have a unknown

145
00:05:48,150 --> 00:05:47,120
basal regulatory role uh there are other

146
00:05:51,350 --> 00:05:48,160
researchers and

147
00:05:51,990 --> 00:05:51,360
our lab is of the opinion that these are

148
00:05:54,710 --> 00:05:52,000
molecular

149
00:05:55,029 --> 00:05:54,720
parasites uh and so they are acting as

150
00:06:02,550 --> 00:05:55,039
uh

151
00:06:05,029 --> 00:06:02,560
beneficial it would be

152
00:06:06,070 --> 00:06:05,039
for parasite to make itself impossible

153
00:06:09,510 --> 00:06:06,080
to be lost

154
00:06:12,950 --> 00:06:09,520
uh for example were this were a

155
00:06:16,870 --> 00:06:12,960
dna into a dnae in teen containing

156
00:06:18,710 --> 00:06:16,880
genome um were to lose its dna e antenes

157
00:06:21,110 --> 00:06:18,720
well now it would have a non-functional

158
00:06:22,950 --> 00:06:21,120
dna e subunit

159
00:06:24,309 --> 00:06:22,960
so as we see in the figure on the on the

160
00:06:26,150 --> 00:06:24,319
left when we have these

161
00:06:27,990 --> 00:06:26,160
if we lose the intent then we have these

162
00:06:29,909 --> 00:06:28,000
two uh

163
00:06:30,950 --> 00:06:29,919

you know useless polypeptide chains that

164
00:06:32,870 --> 00:06:30,960
aren't going to make a functional

165
00:06:34,710 --> 00:06:32,880
protein

166
00:06:36,550 --> 00:06:34,720
and so as i said before these intense

167
00:06:38,150 --> 00:06:36,560
are usually found in highly conserved

168
00:06:38,710 --> 00:06:38,160
housekeeping genes means that it would

169
00:06:41,430 --> 00:06:38,720
be

170
00:06:42,870 --> 00:06:41,440
hard for a cell to lose uh genes would

171
00:06:45,749 --> 00:06:42,880
be hard for a cell

172
00:06:47,670 --> 00:06:45,759
to have these mutate or to mutate or to

173
00:06:50,790 --> 00:06:47,680
have this section of dna of their

174
00:06:52,550 --> 00:06:50,800
of uh sections of these genes lost uh

175
00:06:54,870 --> 00:06:52,560
for whatever reason so

176
00:06:56,309 --> 00:06:54,880
dnae again a very important housekeeping

177
00:06:58,790 --> 00:06:56,319
protein is a subunit

178
00:07:00,390 --> 00:06:58,800
of dna polymerase it's pointed to in red

179
00:07:03,749 --> 00:07:00,400
this is of course involved in

180
00:07:06,070 --> 00:07:03,759
uh the construction of dna uh and

181
00:07:08,070 --> 00:07:06,080
so it cannot function uh without it it

182
00:07:10,390 --> 00:07:08,080
forms the active site we see in the red

183
00:07:13,110 --> 00:07:10,400
circle in the image on the right

184
00:07:13,830 --> 00:07:13,120
with the dna strand double helix we can

185
00:07:16,870 --> 00:07:13,840
see in the middle

186
00:07:17,749 --> 00:07:16,880
buff in blue and then the dna e subunit

187
00:07:22,390 --> 00:07:17,759
highlighted

188
00:07:28,150 --> 00:07:25,589

so if this dna e split intent

189
00:07:29,510 --> 00:07:28,160
was in fact a molecular parasite and

190
00:07:33,350 --> 00:07:29,520
does not play a

191
00:07:35,270 --> 00:07:33,360
functional role uh in terms of its host

192
00:07:36,950 --> 00:07:35,280
i would expect that to be reflected in

193
00:07:39,670 --> 00:07:36,960
its level of distribution

194
00:07:44,869 --> 00:07:39,680
and conservation uh within the

195
00:07:47,430 --> 00:07:44,879
cyanobacteria phylogeny

196
00:07:48,909 --> 00:07:47,440
so here we see my current working

197
00:07:52,790 --> 00:07:48,919
phylogeny of uh

198
00:07:54,710 --> 00:07:52,800
cyanobacteria uh created with a 482 core

199
00:07:55,830 --> 00:07:54,720
proteins found using a program using get

200
00:07:59,510 --> 00:07:55,840
homologs

201
00:08:01,670 --> 00:07:59,520
consisting of a 133 complete genomes

202
00:08:03,830 --> 00:08:01,680
complete high quality genomes off ncbi

203
00:08:05,749 --> 00:08:03,840
with all cyanobacterial families

204
00:08:07,510 --> 00:08:05,759
uh represented in this tree constructed

205
00:08:10,790 --> 00:08:07,520
using iq tree

206
00:08:11,510 --> 00:08:10,800
so this dna e in team is found in about

207
00:08:14,790 --> 00:08:11,520
80

208
00:08:16,550 --> 00:08:14,800
of cyanobacteria genomes and so i first

209
00:08:18,550 --> 00:08:16,560
wanted to pay special attention

210
00:08:21,189 --> 00:08:18,560
into this clade we have highlighted in

211
00:08:26,230 --> 00:08:21,199
this strawberry and cream color

212
00:08:28,550 --> 00:08:26,240
down at the bottom so this clade uh

213
00:08:29,670 --> 00:08:28,560

consists of members of prochlorococcus

214
00:08:33,029 --> 00:08:29,680
seneca caucus

215
00:08:36,310 --> 00:08:33,039
and synovium uh this is often um

216
00:08:38,550 --> 00:08:36,320
widely accepted to be early branching uh

217
00:08:40,949 --> 00:08:38,560
or early or an early branch enclaves off

218
00:08:42,550 --> 00:08:40,959
the cyanobacteria phylogeny

219
00:08:46,310 --> 00:08:42,560
they are related to each other and most

220
00:08:52,550 --> 00:08:49,550
they are all cosmopolitan marine

221
00:08:55,070 --> 00:08:52,560
cyanobacteria and then

222
00:08:56,389 --> 00:08:55,080
in most literature it is referenced that

223
00:09:00,070 --> 00:08:56,399
prochlorococcus

224
00:09:02,949 --> 00:09:00,080
uh does not have any dna in teen

225
00:09:03,590 --> 00:09:02,959
uh in in any of its representative

226
00:09:06,230 --> 00:09:03,600
genomes

227
00:09:06,870 --> 00:09:06,240
uh there are several uh senegal caucus

228
00:09:13,910 --> 00:09:06,880
uh

229
00:09:17,430 --> 00:09:15,190
and then so i thought that would be a

230
00:09:19,590 --> 00:09:17,440
good first subject to uh to delve more

231
00:09:24,870 --> 00:09:19,600
into with the distribution of uh

232
00:09:26,310 --> 00:09:24,880
dna in teens within these genomes

233
00:09:28,630 --> 00:09:26,320
so i'd first like to show you this

234
00:09:33,670 --> 00:09:28,640
nucleotide based phylogeny

235
00:09:38,550 --> 00:09:33,680
based off dna e in

236
00:09:41,990 --> 00:09:38,560
black we see genomes that contain

237
00:09:45,910 --> 00:09:42,000
the dna e split in ten in red we see

238
00:09:47,670 --> 00:09:45,920

genomes that do not contain the intense

239
00:09:49,110 --> 00:09:47,680
and we do sort of get these interesting

240
00:09:52,310 --> 00:09:49,120
clusters uh

241
00:09:54,150 --> 00:09:52,320
as we'd expect we see uh most of the

242
00:09:55,350 --> 00:09:54,160
chlor clock is clustered together that

243
00:09:56,949 --> 00:09:55,360
do not have the in teen

244
00:09:58,550 --> 00:09:56,959
uh we see the senegal caucus and

245
00:10:00,389 --> 00:09:58,560
synovium clustering together

246
00:10:02,949 --> 00:10:00,399
that do not have the in team but we have

247
00:10:06,550 --> 00:10:02,959
a a few interesting outliers the most

248
00:10:09,110 --> 00:10:06,560
notably center caucus wh-8101

249
00:10:10,310 --> 00:10:09,120
uh that does not that does have the

250
00:10:13,430 --> 00:10:10,320
intent but it's

251
00:10:15,910 --> 00:10:13,440
matches very closely uh but it is in the

252
00:10:18,750 --> 00:10:15,920
middle of a of a non-incontaining group

253
00:10:19,910 --> 00:10:18,760
uh and the same with cengage caucus

254
00:10:21,910 --> 00:10:19,920
rs9907

255
00:10:26,630 --> 00:10:21,920
at the bottom so some interesting

256
00:10:30,790 --> 00:10:28,069
and then if we look at the amino

257
00:10:31,269 --> 00:10:30,800
acid-based phylogeny of dna so this is

258
00:10:46,790 --> 00:10:31,279
the

259
00:10:48,150 --> 00:10:46,800
in red and we see a nicer level of

260
00:10:50,230 --> 00:10:48,160
clustering although it is very

261
00:10:53,670 --> 00:10:50,240
interesting how interspersed it gets

262
00:10:57,509 --> 00:10:53,680
early on uh where we'll still see

263
00:11:00,790 --> 00:10:57,519

uh some synecococcus and prochlor caucus

264
00:11:02,630 --> 00:11:00,800
uh uh a matching uh very

265
00:11:04,550 --> 00:11:02,640
very tightly together that do not have

266
00:11:08,069 --> 00:11:04,560
the in teen uh but still

267
00:11:10,710 --> 00:11:08,079
interesting 9907 matches very closely uh

268
00:11:12,630 --> 00:11:10,720
with non-antenna containing uh proteins

269
00:11:18,470 --> 00:11:12,640
but it still has

270
00:11:20,069 --> 00:11:18,480
uh uh that in ten again top and black

271
00:11:21,990 --> 00:11:20,079
but a little bit more tightly clustered

272
00:11:26,829 --> 00:11:22,000
together than

273
00:11:33,030 --> 00:11:29,509
phylogeny

274
00:11:34,870 --> 00:11:33,040
so i would like to uh in the future and

275
00:11:37,350 --> 00:11:34,880
i am currently working on improving

276
00:11:38,550 --> 00:11:37,360
uh the cyanobacteria phytogeny i'd like

277
00:11:41,110 --> 00:11:38,560
to tighten out the nut

278
00:11:42,949 --> 00:11:41,120
tighten down the number of uh homologs

279
00:11:46,069 --> 00:11:42,959
and basing that phylogeny on

280
00:11:48,870 --> 00:11:46,079
uh removing uh genes that are prone to

281
00:11:49,350 --> 00:11:48,880
horizontal gene transfer hgt in order to

282
00:11:51,829 --> 00:11:49,360
uh

283
00:11:53,030 --> 00:11:51,839
get a tighter resolution and get higher

284
00:11:55,990 --> 00:11:53,040
bootstrap values

285
00:11:57,670 --> 00:11:56,000
uh for each of my branches i would like

286
00:11:59,509 --> 00:11:57,680
to continue search for lineages where in

287
00:12:01,110 --> 00:11:59,519
teams are lost and gained

288
00:12:03,750 --> 00:12:01,120

uh or if they could possibly be gained

289
00:12:05,430 --> 00:12:03,760
through uh hgt

290
00:12:06,870 --> 00:12:05,440
i would like to look for further

291
00:12:09,110 --> 00:12:06,880
lineages um

292
00:12:12,069 --> 00:12:09,120
similar to the ones we just highlighted

293
00:12:14,470 --> 00:12:12,079
uh where in teens are

294
00:12:15,350 --> 00:12:14,480
either uh we're we're in things that are

295
00:12:17,030 --> 00:12:15,360
either present

296
00:12:18,870 --> 00:12:17,040
where we have a mixture of uh in team

297
00:12:21,750 --> 00:12:18,880
containing and non-incontaining

298
00:12:23,590 --> 00:12:21,760
of related genomes and i would like to

299
00:12:27,110 --> 00:12:23,600
compare the phylogeny of

300
00:12:28,870 --> 00:12:27,120
dna extines to the dna in teams uh

301
00:12:31,190 --> 00:12:28,880
in in the consensus for dianobacteria

302
00:12:33,829 --> 00:12:31,200
phylogeny uh to see if we see uh

303
00:12:39,269 --> 00:12:33,839
a conservation of in teen sequences uh

304
00:12:41,430 --> 00:12:39,279
or not

305
00:12:44,389 --> 00:12:41,440
so thank you very much uh for listening

306
00:12:47,590 --> 00:12:45,910
i would additionally like to thank and

307
00:12:49,430 --> 00:12:47,600
acknowledge all my committee

308
00:12:51,269 --> 00:12:49,440
members all the members of the go garden

309
00:12:53,670 --> 00:12:51,279
lab and especially