**BioBERT** is a pre-trained language model that is specifically designed for biomedical text mining tasks. It is an extension of the BERT (Bidirectional Encoder Representations from Transformers) model, which is a widely used natural language processing model. BioBERT is trained on a large corpus of biomedical literature, such as PubMed, and clinical notes, such as MIMIC-III.

Detailed description of the pretraining process for BioBERT:

1. **Data Collection**: The first step in pretraining BioBERT is to collect a large corpus of biomedical text. In the case of BioBERT, this includes PubMed abstracts and full-text articles, as well as clinical notes from the MIMIC-III database.
2. **Text Cleaning**: Once the data is collected, it is cleaned to remove any personally identifiable information, such as patient names, and other sensitive information that may violate privacy laws. The text is also processed to handle biomedical abbreviations and entities, such as chemical names, gene names, and protein names.
3. **Tokenization**: The text is tokenized using a variant of the WordPiece tokenizer, which is used in the original BERT model. However, BioBERT uses a vocabulary that is specifically designed for biomedical text, which includes words and subwords that are specific to the biomedical domain.
4. **Masked Language Modeling**: The pretraining objective used in BioBERT is masked language modeling (MLM). In MLM, a certain percentage of tokens in the input text are masked, and the model is trained to predict the masked tokens based on the surrounding context. For example, if the input text is "The patient was diagnosed with [MASK] cancer", the model is trained to predict the masked word, which is "lung" in this case.
5. **Next Sentence Prediction**: In addition to MLM, BioBERT also uses a next sentence prediction (NSP) task. In NSP, the model is trained to predict whether two sentences in the input text are consecutive in the original text or not. This helps the model to understand the relationships between sentences in a document.
6. **Training**: The pretraining is done using a variant of the BERT model architecture, which uses a transformer-based neural network to process the input text. The model is trained using a large amount of computational resources, such as graphics processing units (GPUs), to speed up the training process.
7. **Fine-Tuning**: After pretraining, the BioBERT model can be fine-tuned on specific downstream tasks, such as named entity recognition or relation extraction, to improve its performance on those tasks. This involves training the model on a smaller dataset that is specific to the downstream task.

In summary, the pretraining process for BioBERT involves collecting a large corpus of biomedical text, cleaning and tokenizing the text, training the model using MLM and NSP objectives, and fine-tuning the model on specific downstream tasks.